DOI: 10.51845.37.1.7

Evolution Is Neither Random Accidents nor Divine Intervention: Biological Action Changes Genomes

by James A. Shapiro

Charles Darwin and his followers postulated that random accidental mutations of small effect plus natural selection over long periods would provide sufficient hereditary variation to explain biological diversity. Research since the middle of the twentieth century has unexpectedly shown that living organisms possess many different means of altering their genomes biologically, and these processes have been validated by DNA sequence analysis. In addition, the biological process of interspecific hybridization has become recognized as a major source of rapid speciation and genome amplification. Thus, it is time to shift our basic concept of evolutionary variation from the traditional model of slow change from non-biological sources to a fully biological model of rapid genome reorganization stimulated by challenges to reproduction.

Introduction

In Western society prior to the Enlightenment, there was little disagreement about the origins of biological diversity: it resulted from divine creation of an unchanging panorama of plant and animal species, as explained in *Genesis*. No thought was given to the idea that living organisms could change their fundamental natures. Even a scientist dedicated to analyzing the nature and classification of life forms, Carl Linnaeus (1707-1778), and one who documented the extinction of fossil organisms, Georges Cuvier (1769-1832), both believed in the fixity of species.

The first naturalists to write about the evolutionary origins of biological

diversity through "descent with variation" from natural causes were Erasmus Darwin, Charles's grandfather (in *Zoonomia; or, The Laws of Organic Life*, 1794) and Jean-Baptiste Lamarck (in *Philosophie Zoologique,* 1809) at the turn of the nineteenth century. In the middle of the nineteenth century, Charles Darwin (*On the Origin of Species by Means of Natural Selection*, 1859) and Alfred Russell Wallace (*On the Tendency of Varieties to Depart Indefinitely From the Original Type*, 1858) expanded the argument for natural evolutionary transformations. Whereas Lamarck had postulated an undefined "*pouvoir biologique*" ("life power") underlying and directing hereditary changes, Darwin and his neo-Darwinist followers avoided any implication of biological purpose or action in heritable variation.

Today, the mainstream neo-Darwinist school insists that hereditary changes result from undirected random accidents that inevitably arise in organismal reproduction. This is consistent with Darwin's own emphasis on evolution as slow change over many generations, guided purely by natural selection (phyletic gradualism). In this, he probably was influenced by the Uniformitarian philosophy espoused by his Edinburgh geology professor, Charles Lyell. Darwin even proposed a test for his postulate that long sequences of changes of small effect were the sole determinants of evolutionary trajectories: "If it could be demonstrated that any complex organ existed which could not possibly have been formed by numerous, successive, slight modifications, my theory would absolutely break down. But I can find out no such case."

Over a century later Ernst Mayr, a leader of the mid-twentieth century neo-Darwinist "Modern Synthesis," wrote: "The proponents of the synthetic theory maintain that all evolution is due to the accumulation of small genetic changes, guided by natural selection and that trans-specific evolution [i.e. origins of new species and taxonomic groups] is nothing but an extrapolation and magnification of the events that take place within populations and species."

The dispute between religious and naturalistic accounts of species origins is ongoing. The conflict persists to the present day among a significant fraction of the U.S. population, and there are serious movements to ban the teaching of evolution in schools. Support for evolution guided by divine intervention has a toehold in the quasi-scientific Intelligent Design (ID) movement, initiated by Michael Behe (*Darwin's Black Box: The Biochemical Challenge to Evolution*, 1996) and carried on by members of the Discovery Institute and other creationist think tanks. The basic argument that ID theorists make is that natural selection of random hereditary changes cannot produce genomes capable of expressing all the intricate networked adaptations modern molecular biology has revealed to operate in living organisms. This conundrum is, in Behe's words, "irreducible complexity." Hence, the ID theorists posit a need for divine intervention.

The ID argument has a valid point with regard to the explanatory limits of

neo-Darwinism, still widely regarded as the only legitimate scientific explanation of evolution. ID falls down by assuming (as do mainstream evolutionists) that genome change occurs from outside the boundaries of life itself. Within the scientific community, there is agreement that the hereditary variation necessary for evolutionary change occurs by natural means. But significant difference exists between scientists about what constitutes "natural means."

While random mutation leading to gradual change (phyletic gradualism) was a reasonable assumption to make in 1859 and even in the 1940s (when the Modern Synthesis was proposed), the scientific understanding of how genome change actually occurs has grown tremendously since then. There have been a series of revolutionary changes in our analysis and understanding of heredity and the processes behind genome evolution.

I. Barbara McClintock's discoveries of active chromosome break repair, and later, of mobile genetic elements capable of migrating to new places in the genome was a critical development. These amount to changes in syntax: a mobile genetic element may induce genetic instabilities at its landing site, for example.

McClintock's studies began in the 1920s, characterizing X-ray-induced mutants of maize, originally thought to carry ordinary "gene mutations." Her cytological investigations showed they were in fact the results of cells rearranging their genomes by joining the broken ends of chromosomes damaged by the radiation. This totally unexpected result revealed that living organisms possess biological systems for rapid and non-random genome change.

Later, in the 1940s, McClintock studied maize plants she engineered to have chromosomes that broke at every cell division from fertilization on. Without intending it, she created a "genomic earthquake" that resulted in a discovery that transformed our understanding of hereditary variation. She showed that genomes contain components capable of changing their chromosomal location (transposing). This contradicted mid-twentieth century beliefs that genes were integral units occupying fixed positions in the chromosome. McClintock named her mobile genetic elements "controlling elements" because they altered developmental patterns of genome expression from nearby fixed loci. By 1951, McClintock could see that her results indicated a more advanced and revolutionary idea of genome organization, distinguishing between the specific functional character coding content of a genetic locus (e.g. eye color, sugar metabolism, limb structure, etc.) and the associated controlling elements that regulate its expression. This kind of thinking was so novel that it is not surprising her initial presentation at the 1951 Cold Spring Harbor symposium was greeted with incomprehension and hostility.

Today, we recognize controlling elements in maize as the first examples of a vast range of mobile genetic elements present in all organisms from bacteria to plants and animals. The movement of a

transposable element to a new location has the potential to alter the signals regulating expression of a nearby coding sequence: its control module, in a phrase. For evolutionary variation, regulation of genome expression altered by mobile DNA has at least three distinct implications:

- As McClintock discovered, insertion of a transposable element next to a genetic locus can confer novel regulation on expression of its coding sequences. This evolutionary process has been confirmed for thousands of genetic loci analyzed in fully sequenced genomes.
- The changes are not random accidents but are *biological* in nature. That means they involve the action of defined genome components subject to regulatory systems that control when and how frequently they transpose, as well as their target specificity.
- Since transposable elements can insert at multiple locations, they can establish coordinated genomic networks by inserting the same control signals at each network's component genetic loci. In this way, transposable elements help answer the Intelligent Design critique about the impossibility of naturally evolving irreducible complexity.

Such flexibility and natural biological control over the timing and outcomes of evolutionary variation were literally inconceivable to the mid-twentieth century founders of the neo-Darwinist Modern Synthesis. Nonetheless, today they are established science and essential to a contemporary understanding of how evolution works.

II. The identification of DNA as the physical basis of genome coding in 1953 ultimately made it possible to read genome sequences and precisely define the DNA differences between all manner of hereditary variants and between distinct species.

Even before genome sequencing became possible in 1968, Roy Britten and David Kohne discovered a key distinction between the genomes of prokaryotic bacteria (which do not contain their genomes in a cell nucleus) and those of complex eukaryotes like plants and animals (which contain their genomes within a nucleus). The prokaryotic genomes contained almost exclusively unique DNA sequences, for example, while the genomes of humans and other complex eukaryotes contained significant fractions of repetitive DNA sequences. At the time, only unique or slightly repeated sequences were considered to carry significant genetic information. As a consequence the repetitive DNA was labelled as "junk DNA," "selfish DNA," or "selfish genetic elements." Richard Dawkins famously erected a widely popular philosophy of evolution on the basis of "*The Selfish Gene*" (1976).

Today, we recognize that most of this repetitive DNA is made up of transposable elements and other repeats needed for various aspects of genome function, especially developmental regulatory networks controlling cellular differentiation. The repeats help guide the origin

of cell lines that comprise distinctive tissues, say bone tissue versus nervous tissue. Both have the same DNA, yet each cell type expresses the genome in distinctive ways controlled by different DNA repeats.

The conventional conception of the genome as an assemblage of function-specifying genes thus falls short. When the first version of the human genome was published, for example, it was found to contain over 3,000,000 copies of various kinds of transposable elements comprising at least 45 percent of the entire genome. By comparison, the putatively all-important protein-coding genes were estimated at 30,000, less than 1.5 percent of the genome. This unexpected disjunction came to be known as the genome size or "C-value Paradox." When the protein-coding and non-coding contents of several genomes were plotted against organismal complexity (judged by number of cell types), the protein-coding DNA peaked and levelled off at 10^7 - 10^8 nucleotide base-pairs. In contrast, the non-coding content continued to increase logarithmically to between 10^9 and 10^{10} nucleotide base-pairs. Furthermore, non-coding and repetitive DNA is the most evolutionarily "volatile" component of genomes. In any particular taxonomic group, the vast majority of proteins are shared, but there can be dramatic changes in the mobile repetitive elements between closely related "sibling species." Clearly, to account for these unexpected observations, something is missing from our conventional understanding not only of genome evolution but even of genome function.

Conventional evolutionary theory deals largely with protein evolution, and there have been several surprises there as well. DNA sequencing became possible in 1977. The first unexpected discovery was that protein-coding sequences in many eukaryotes and their viruses were not continuous regions of the genome. Rather, the sequences encoding parts of the protein could be separated by intervening stretches of DNA that did not encode any part of the protein. The partial coding sequences were dubbed "exons" (*ex*pressed elements) and the non-coding regions between exons were labelled "introns" (*in*tervening elements). The entire discontinuous coding region of exons and introns could be transcribed into RNA, following which the segments corresponding to the introns had to be "spliced" out of the primary RNA to form a translatable messenger mRNA that could encode the final protein product. Special "splice donor" and "splice acceptor" sequences flank the introns to guide the splicing process. The result is a kind of genetic grammar: a genetic locus transcript with multiple introns can be spliced in various ways to encode related but distinct proteins, ending the concept of genes as unitary entities. Evolution has used such "alternative splicing" to diversify and fine-tune the expression of genetic loci for functions that range from expanding protein-protein interactions, to sex determination in mice, to stress responses in fighting fish.

Alternative splicing also highlights the unexpected Lego-like organization and evolution of many proteins that can quickly gain or lose part of their polypeptide sequences. Comparative analysis revealed that many proteins are linear composites of multiple structurally defined polypeptide regions called "domains." Each domain executes a defined molecular task (e.g. DNA binding, protein-protein interaction, insertion into cell membranes, cleavage of a specific covalent bond) that becomes part of the overall protein functionality. Domains were initially identified when genome sequence comparisons revealed that many functionally distinct proteins contained virtually identical regions in combination with otherwise divergent sequences. It became apparent that proteins can evolve combinatorially and efficiently by exon shuffling or insertion into novel coding contexts, not just by collecting random single amino acid changes as envisioned by the conventional model.

Not only can exons duplicate and move from one genetic locus to another within the same genome, entire protein-coding DNA segments can be transferred from one species to another across taxonomic boundaries in a process known as horizontal gene transfer (HGT). Sequence analysis identifies horizontal transfers when a novel protein abruptly appears in a phylogeny with clear ancestry from the same type of protein in a taxonomically distant organism. For example, herbivory emerged rapidly among both nematode worms and beetles through the acquisition of enzymes from diverse bacteria and fungi that are capable of digesting complex plant polymers.

HGT illustrates the principle that no genome is completely isolated from other genomes. There are several possible ways such transfers may occur, but all depend upon biological functions, and none of them involve random mutations. A couple of paths for HGT involve viruses acting as gene carriers, reminding us of the potent effects these peripatetic biosphere inhabitants can play in organismal evolution. Viruses can integrate in cellular genomes, where they can transport DNA encoding important functions, such as toxin production, as happened in major bacterial pathogens. In mammalian evolution, endogenized retroviruses constitute a major class of transposable elements that facilitated formation of the placenta, pluripotency of stem cells, pre-implantation embryonic development, innate immunity, and other vital functions.

Although not widely recognized, both genome sequencing and real time experiments have revealed a purely biological trigger for speciation. Reading genome sequences has confirmed that crosses between closely related species and whole genome duplications (WGDs) have played major roles in evolution. Genome sequencing frequently reveals what are called "introgressions" when the sequence of one species' genome has slightly different regions that are the same as those in a closely related but distinct "sibling" species. This kind of

observation indicates that the two species had inter-bred in the past.

This is significant because the most effective practical method we currently have to initiate the formation of new species is to mate (or hybridize) sibling species. Hybrid speciation has been how many of our cultivated crops originated (wheat, oats, cotton, rapeseed, etc.). It has also been observed in the wild, for example, with the speciose Darwin's finches. A recent paper shows how repeated cycles of interspecific hybridization have led to extensive speciation and tremendous phenotypic diversification among the cichlid fishes of Lake Tanzania. It seems likely that interspecific hybridization occurs most often when intraspecific mating populations decline (and where novel organismal capabilities are most beneficial), establishing an adaptive connection between cause and effect.

Interspecific hybridization frequently produces progeny with tremendous genome instability, involving activation of transposable elements and chromosome rearrangements. When the hybrids reproduce, they can become progenitors of totally new species with new traits and new genome configurations in just a few generations. Thus, long periods of selection are not essential to taxonomic divergence. In self-pollinating plants in particular, but also in animals, hybrid speciation is often accompanied by whole genome duplication events, which stabilize the genome and increases fertility. It also expands the DNA substrate available for further evolutionary development of new functionalities because one copy of each locus can be repurposed without endangering existing functions encoded by another copy. The evolutionary history of eukaryotes, ranging from yeast and fungi to flowering plants and animals, is marked by a succession of WGD events. It is difficult to imagine a process further from random mutations than WGD, which involves control of complex cell cycle and nuclear division processes. It is not hard to see repeated doubling in genome coding capacity as one source of greater organismal complexity with ongoing evolution.

III. Molecular analysis of genome change systems documents the many different ways that biochemical and cellular functions alter the content of cellular genomes.

The discovery of DNA as the genetic coding medium in 1953 triggered the enormous research undertaking known as molecular biology. One major facet has been replication, recombination, repair, and restructuring of the genome, with special emphasis on elucidating how genetic variability arises at the level of DNA. This has revealed an unexpectedly wide range of intrinsic biochemical processes that produce different types of highly non-random genome change. For example, all living cells have the biochemical tools to cut and splice DNA molecules. In technical terms, cells have proteins that open and ligate (stitch together) the phosphodiester bonds in DNA strands. I call this capacity natural genetic engineering (NGE). Various pro-

teins carry out these operations with a whole range of specificities, including chromosome rearrangements, transposition, exon shuffling, and integration of horizontally acquired DNA. Other mechanisms allowing for fast non-random genome change include localized strings of mutation ("kataegis" or thunderstorms) and somatic hypermutation; transposons and "retrotransposons" which relocate DNA sequences; Double-Strand-Break repair which often generates complex sequence aggregates; and Chromoanagenesis, an umbrella term for a class of rapid multisite chromosome rearrangements.

In short, these mechanisms show that long periods of time are not required to accumulate extensive levels of genome restructuring for evolutionary diversification. The complexities emphasize the inherently *biological* nature of complex evolutionary variation, coming as they do from cellular activity.

IV. Molecular analysis of how genomes function as databases has revealed that genomes do not obey the limits of Crick's Central Dogma of Molecular Biology. Instead of coding principally for proteins (via messenger RNAs), genomes also comprise a broad range of active non-coding RNAs (ncRNAs).

The significance of ncRNAs was revealed by the ENCODE (*Enc*yclopedia of *D*NA *E*lements) project (an offshoot of the Human Genome Project), which attempted to comprehensively characterize the significance of all DNA sequences in the human genome. ENCODE documented pervasive, cell type-specific transcription of *all* classes of genomic DNA, not just the protein-coding regions. In other words, transcription of DNA into RNA produces not just protein-coding messenger RNA, but abundant ncRNAs. These are frequently transcribed from transposable element DNA.

We now know that many ncRNAs have intricate structures that play critical roles in biochemical function, including: processing other RNA transcripts; providing scaffolds for the assembly of multimolecular complexes; nucleating biomolecular condensates; targeting epigenetic modifications across the genome; regulating genome expression; and regulating cellular differentiation, especially higher nervous system development. For example, one genome-wide association study (GWAS) on human higher cognitive function (using test score data) identified 267 genome regions that are exclusive to humans and not found in other primates. The vast majority (95 percent) of these regions encode an ncRNA rather than a protein.

By demonstrating that all regions of the human genome are templates for the synthesis of functional biomolecules, the discovery of ncRNA resolved the C-value paradox (the unexpectedly small component of the genome that encodes proteins). This major enhancement in our understanding of genome database contents has been thoroughly described in a 2023 book by John Mattick and Pablo Amaral: *RNA, the Epicenter of Genetic Information: A new understanding of molecular biology*, which is available as a free download online.

The ncRNAs add another layer of complexity to multicellular regulatory circuits on top of the protein-based networks familiar to us. Many of them affect high level decisions in embryonic development, which are likely to influence characters such as morphology, coloration, sensory processing and behavior that frequently distinguish new species from their progenitors, with whom they will generally share basic metabolic and structural features typical of their genus and higher taxonomic groupings.

A New Evolutionary Paradigm

The foregoing trends in contemporary genetic and evolutionary studies lead to important conclusions. Since the formulation of the neo-Darwinian Modern Synthesis in the 1940s, many unexpected surprises have emerged from genetics research and from novel molecular techniques that define evolutionary genome changes. These include:

1. mobile genetic controlling elements;
2. discontinuous coding of proteins;
3. protein evolution by domain/exon swapping;
4. horizontal DNA transfers between unrelated taxa;
5. complex eukaryotic cells have molecular and cellular processes for major genome restructuring and new sequence creation;
6. there is widespread rapid speciation by interspecific hybridization;
7. ncRNAs encoded partly by repetitive DNA elements fill major regulatory roles in cell and developmental biology, and;
8. DNA that codes for ncRNA instead of proteins comprises the majority of genomes in the most complex organisms.

Surely, the knowledge of all these novel genome features justifies a reappraisal of our fundamental assumptions about genetics and how we envision evolutionary change. From these multiple sources of evolutionary variation, we can now envision a new, more biological and functional picture of the genome evolutionary process. Rather than natural selection acting over long periods of time on small random modifications of multiple independent phenotypic traits to produce diverse life forms, the biological paradigm posits that organisms possess inherent capacities for rapid concerted genomic innovations to evolve when species survival is endangered. One trigger for these innovations can be interspecific hybridization, which will increase in frequency as mating pools shrink because of adverse conditions. Whole Genome Duplications (WGDs) accompanying interspecific hybridization will also increase genomic capacity as evolution proceeds, providing the opportunities for evolving novel proteins and ncRNAs to increase organismal complexity. In other words, biology has re-entered the evolutionary process in transformative ways, largely documented by molecular evidence.

James A. Shapiro *is a professor in the Department of Biochemistry and Molecular Biology, University of Chicago, emeritus.*