Inverse relationship between genetic diversity and epigenetic complexity

Shi Huang

The Burnham Institute for Medical Research
10901 North Torrey Pines Roads
La Jolla, CA 92037
shuangtheman at yahoo.com

Key words: Genetic equidistance result, evolution, molecular clock, epigenetic complexity, maximum genetic diversity hypothesis, Darwinism

Abstract Early studies of molecular evolution revealed a correlation between genetic distance and time of species divergence. This observation provoked the molecular clock hypothesis and in turn the 'Neutral Theory', which however remains an incomplete explanation since it predicts a constant mutation rate per generation whereas empirical evidence suggests a constant rate per year. Data inconsistent with the molecular clock hypothesis have steadily accumulated in recent years that show no correlation between genetic distance and time of divergence. It has therefore become a challenge to find a testable idea that can reconcile the seemingly conflicting data sets. Here, an inverse relationship between genetic diversity and epigenetic complexity was deduced from a simple intuition in building complex systems. Genetic diversity, i.e., genetic distance or dissimilarity in DNA or protein sequences between individuals or species, is restricted by the complexity of epigenetic programs. This inverse relationship logically deduces the maximum genetic diversity (MGD) hypothesis, which suggests that macroevolution from simple to complex organisms involves a punctuational increase in epigenetic complexity that in turn causes a punctuational loss in genetic diversity. The hypothesis fully grants Neo-Darwinism to be what it really is (a theory of microevolution) and explains all the major facts of evolution. Importantly, it predicts the most remarkable result of molecular evolution, the genetic equidistance result, which originally provoked the molecular clock hypothesis.

> The only real valuable thing is intuition……The whole thing of science is nothing more than a refinement of everyday thinking.
>
> - Albert Einstein

> All sciences are nothing but human wisdom…… The two operations of our understanding, intuition and deduction, on which alone we have said we must rely in the acquisition of knowledge.
>
> - Rene Descartes

Introduction

It is remarkable that the human mind is able to comprehend nature. The scientific understanding of nature is largely based on mathematics. Since mathematics is premised on axioms or self-evident intuitions, it can be easily inferred that intuition is the ultimate foundation of science. The relationship between intuition and a natural phenomenon is sometimes indirect or follows the hierarchy from intuition to mathematics, to physics, to chemistry, and to biology. But it can also be direct, for example, Newton's three laws of motion were originally postulated as axioms. However, an intuition-based law or axiom of biology has yet to be uncovered. Biology will not become a true science on a par with mathematics or physics until it can evolve from a positive science, describing how things are, into a normative one, telling nature how things should be. An intuition-based theory is true on its own logical coherence (like a mathematical proof) and does not in principle need validation from empirical data. In contrast, no amount of experimental data could prove a provisional theory that is based on observations. And a single exception is sufficient to doom such a theory regardless how many supporting data it may have. A truly scientific theory must not allow any exceptions within its domain of application, because once it does, it automatically renders itself non-testable or makes testing meaningless, and would no longer have any predictive value or qualify as scientific.

The Neo-Darwinian theory is the dominant mainstream theory for evolution and widely taught to biologists and the public at large. However, it is not possible to use this theory to explain the major facts of molecular evolution. Its ad hoc substitute for the domain of molecular evolution, the molecular clock hypothesis, is essentially unknown outside the circle of evolution specialists. This hypothesis must negate the idea of selection, the cornerstone of Neo-Darwinism. The co-existence of two vastly different and non-connected theories to account for two different but inseparable aspects of evolution, phenotype versus genotype, is plain evidence that neither is a complete theory of evolution. Indeed, all existing theories of evolution have numerous factual contradictions and take exceptions for granted. Therefore, there should exist a better theory that can explain all major facts of evolution via a single universal theme.

The molecular clock hypothesis was triggered by the empirical observation of a correlation between genetic distance as measured by DNA or protein sequence dissimilarity and time of species divergence as inferred from fossil records. In the early days of molecular evolution studies, genetic distance was simply represented by percent identity in a given protein sequence. Two kinds of sequence alignment can be made using the same set of sequence data. The first aligns a recently evolved organism such as a mammal against those simpler or less complex species that

evolved earlier such as amphibians and fishes. The second aligns a simpler outgroup organism such as fishes against those more complex sister species that appeared later such as amphibians and mammals.

The first alignment indicates a near linear correlation between genetic distance and time of divergence, implying indirectly a constant mutation rate among different species. For example, human is closer to mouse, less to bird, still less to frog, and least to fish. The second alignment shows the genetic equidistance result where sister species are approximately equidistant to the simpler outgroup. For example, human, mouse, bird, and frog are all equidistant to fish in any given protein dissimilarity. Since all of the sister species are also equidistant in time to the outgroup fish, this directly triggered the idea of constant or similar mutation rate among different species, no matter how different they may be. Since both alignments use the same sequence data set, certain information may be revealed by either alone. But the data that most directly and obviously support the interpretation of a constant mutation rate is the genetic equidistance result.

The molecular clock hypothesis was first informally proposed in 1962 based largely on data from the first alignment [1]. Margoliash in 1963 performed both alignments and made a formal statement of the molecular clock after noticing the genetic equidistance result [2, 3]. "It appears that the number of residue differences between cytochrome c of any two species is mostly conditioned by the time elapsed since the lines of evolution leading to these two species originally diverged. If this is correct, the cytochrome c of all mammals should be equally different from the cytochrome c of all birds. Since fish diverges from the main stem of vertebrate evolution earlier than either birds or mammals, the cytochrome c of both mammals and birds should be equally different from the cytochrome c of fish. Similarly, all vertebrate cytochrome c should be equally different from the yeast protein."

The comparisons that produced the equidistance result, as Margoliash stated [2], "disregard the relation of amino acid substitutions observed to the actual number of effective mutational events which occurred." So, the genetic equidistance result and the molecular clock hypothesis were originally established by percent identity in protein sequences. The actual number of mutational events in the past evolutionary process is irrelevant to the equidistance result, and is impossible to discern anyway if the percent nonidentity in fact represents the maximum that has long been reached before present time.

The genetic equidistance result is extremely robust and universal and has been independently confirmed for numerous proteins and numerous species [4]. It is the most remarkable result of molecular evolution since it was completely unexpected from classical Neo-Darwinian theory. However, what has become popular known today is not the result itself but the molecular clock interpretation of it [5-7].

The molecular clock hypothesis asserts that the rate of amino acid or nucleotide substitution is approximately constant per year over evolutionary time and among different species. Two different species are thought to gradually accumulate mutations over time since their most recent common ancestor. Their genetic distance in ancient times is thought to be smaller than their distance today. None of these assertions are based on intuitions or could be considered as self-evident. Nor do they have direct experimental support. They are all ad hoc interpretations of the genetic equidistance result.

The empirical observation of an apparently constant mutation rate has provoked the 'Neutral Theory'. But this theory is now widely acknowledged to be an incomplete explanation."[8, 9].

The common practice of relative rate tests that often interprets small deviations from an exact equidistance as being statistically significant overlooks the striking fact that the deviations are rarely large. If the real phenomenon here is non-equidistance with equidistance being coincidental, one would expect to see much larger variations in distance. Thus, the data shows that the real phenomenon here is equidistance [4].

Although there clearly exists a correlation between genetic distance and time of divergence, such correlation is not universal and is often violated as more data became known in recent years. Numerous studies based on extant organisms have questioned the constancy of mutation rate [5-14]. The genetic distance between two subpopulations of medaka fish that had diverged for ~ 4 million years is 3-fold greater than that between two different primate species (humans and chimpanzees) that are thought to have diverged for 5-7 million years [15]. The genetic distance measured on genealogical timescales (< 1 million years) is often an order of magnitude greater than that on geological timescales (> 1 million years) [10, 16], suggesting that genetic distance measured in evolutionary time is independent of actual mutation rate measured in real time.

Importantly, few recognized that results of violations of a constant molecular clock do not invalidate the genetic equidistance result [4]. The equidistance result does not necessarily mean rate constancy. The constant mutation rate interpretation of the equidistance result represents an over-interpretation of the actual result, since the result shows merely the outcome of evolution and says nothing about the past mutation process.

A recent study of DNA and protein sequences of ancient fossils (Neanderthals, dinosaurs, and mastodons) challenged a fundamental premise of the modern evolution theory [17]. It shows that genetic distance had not always increased with time in the past history of life on Earth. Neanderthals are more distant than modern humans are to the outgroup chimpanzees in non-neutral DNA sequences, contrary to expectations from the molecular clock interpretation of the genetic equidistance result [17]. This unexpected observation has been independently confirmed by analysis of Neanderthal mitochondrial protein sequences [18].

Given the numerous factual exceptions, it is clear that the molecular clock hypothesis or the Neo-Darwinian theory can not qualify as a true or proven scientific theory at least in the domain of macroevolution. Both theories have already been falsified by numerous tests. A new and more complete idea is needed that must explain all the major facts of evolution and must not have factual contradictions. It also must grant the proven virtues of the existing theories within their specific domain of application. Here, a simple intuition in building complex systems was used to deduce a novel axiom of biology, the inverse relationship between genetic

diversity and epigenetic complexity. This axiom or its logical deduction, the maximum genetic diversity hypothesis, was found to deduce or explain all the major facts of evolution, including both for and against the correlation between genetic distance and time of divergence. Importantly, it predicts the most remarkable result in molecular evolution that originally provoked the constant clock hypothesis, the genetic equidistance result, while fully grants different mutation rates to different species.

An intuition in building complex systems

It is a self-evident intuition that simpler systems or machines can tolerate more variations or choices in building blocks. The more complex the system, the more restriction would be placed on the choice of building blocks. A one-story house can be build by all varieties of bricks but only the stronger ones among them can qualify for a 100-story building because the weaker ones cannot withstand the weight of a 100-story building. The number of choices of different materials for constructing a toy bicycle is much greater than that for a space shuttle. The inverse relationship between building block diversity and system complexity is here termed the first axiom of construction.

Complex organisms and epigenetic programs

The major building blocks for biological organisms are DNAs. The complexity of organisms is reflected by the ways a set of DNAs is used to make a cell or an organism with multiple distinct cell types. The more the cell types, the more the number of ways of using the same set of DNAs, and the more complex the organism [19-26]. Phenotypes are determined by the primary sequence of DNAs or genotypes as well as by the ways by which DNAs are used or expressed, often termed epigenotypes or epigenetic programs. Each cell type represents a distinct epigenetic program of the same genotype. Cell types with distinct functions differ only in epigenotypes but not in genotypes (an extremely small number of special cell types such as antibody producing cells are exceptions).

Epigenetic programs are not only inherited during mitotic cell division but are also transmitted through the germline to the next generation [27-29]. The next generation receives not only genetic information encoded in DNA but also epigenetic information carried by the non-DNA molecules of the fertilized egg.

Epigenetic programs control both expression levels of genes and the specific combination of co-expressed genes within a specific cell type. The epigenetic programs are here broadly defined, including both the primary epigenetic proteins as well as those secondary or tertiary proteins that could regulate the primary proteins. The number of human genes is only about 1.6 fold more than that of a fruit fly and about the same as the mouse or fish. However, the number of certain enzymes responsible for epigenetic gene organization, the PRDM subfamily of histone methyltransferases, increases dramatically during metazoan evolution: 0 in bacteria, yeasts, and plants; 2 in worms, 3 in insects; 7 in sea urchins, 15 in fishes, 16 in rodents, and 17 in primates [30, 31]. Also, the core histone genes H2A, H2B, H3, and H4 have been duplicated in humans but not in chimpanzees [32]. This faster pace of expansion of certain epigenetic enzymes, relative to the pace for the genome, in complex metazoan indicates a correlation between complex epigenetic programs and complex organisms.

In addition, microRNAs are an important part of the epigenetic program, and the number of microRNA genes correlates well with organismal complexity [33, 34]. The relative amount of non-coding sequences increases consistently with complexity [35]. Complex organisms also show complex gene expression patterns as indicated by the fact that 94% of human genes have alternative products or alternative splicing relative to only 10% in *C. elegans* [36, 37].

Complex organisms are here defined as those that have complex epigenetic programs. Whether an organism is more complex than another organism can be roughly estimated based on a comparison of the number of genes involved in epigenetic programs. This is informative to differentiate unicellular organisms: yeasts have more epigenetic enzymes than bacteria and are therefore more complex; yeasts have several histone acetylases and SET domain histone methyltransferases while bacteria have none. Based on the number of the PRDM family of epigenetic enzymes, it is also easy to conclude that vertebrates are more complex in epigenetic programs than invertebrates or that primates are more complex than rodents or fishes.

When the numbers of epigenetic enzymes are similar for some multicellular organisms, then the number of tissue or cell types is a good measure of epigenetic complexity since each tissue or cell type is representative of a distinct epigenetic program or gene expression pattern. The more tissue types an organism has, the more the number of distinct epigenetic programs and hence the more complex the epigenetic program. The exact number of tissue types for any complex organism remains unknown, largely because there are many more neuronal cell types than we can presently recognize [38]. But this may not prevent one from drawing the conclusion that organisms that appeared early in evolution generally have less number of cell types than their descendant but distinctly different organisms that appeared later.

The number of neuronal cell types likely represents a major proportion of the total number of cell types in a complex animal. Also, epigenetic programs may control the complex interaction and organization of these neuronal cell types that manifest as intelligent brain functions. Thus, organisms with complex and intelligent brains are likely to contain more cell types or more complex interaction and organization of neuronal cell types. It is therefore easy to infer that the first primate has more cell types or complex organizations than the first mammal which has more cell types or complex organizations than the first vertebrate. Also, animals that go through complex and prolonged developmental process contain more complex epigenetic programs since the development from a fertilized egg to an adult organism is largely an epigenetic process. The same tissue type often exhibits different expression patterns or epigenetic programs at different stages of development.

Organisms with the most complex and advanced brain (but not necessarily the largest in volume) are necessarily more complex in epigenetic programs or have more varieties of neuronal cell types and more complex interactions. Humans obviously have more distinct cell

types and more complex neuronal interactions, thanks to our complex brain, than any other species that ever lived and are necessarily the most complex and diversified in epigenetic programs. Humans have ~700 billion neurons while mice have only ~70 million. Human brain shows dramatically more methylated DNAs than chimpanzees [39].

Inverse relationship between genetic diversity and epigenetic complexity

From the self-evident intuition of building complex machines, it is easy to deduce an equivalent principle or axiom in constructing biological organisms. Thus, simple organisms with low epigenetic complexity can tolerate more variations in DNA or have higher genetic diversity. There exists an inverse relationship between genetic diversity and epigenetic complexity. Genetic diversity is defined here as genetic distance or dissimilarity in DNA or protein sequences between different individuals or species.

Simple organisms are built more by the primary function of a gene rather than by a specific expression pattern of the gene. A gene may only have one expression pattern in simple organisms and many variants of the gene may be able to fit within that one expression pattern. In contrast, when an organism is built by multiple distinct gene expression patterns or cell types, the variation in gene sequence would be necessarily restricted.

The reason is easy to understand. If cell type A is determined by expression pattern X and cell type B by pattern Y of the same gene, a mutational variant of the gene must be compatible with both expression pattern X and pattern Y. Such multilevel compatibility reduces the number of variants of the gene that can meet the multiple requirements. If ten mutational variants can fit with expression pattern X, then may be only three of the ten would fit with both patterns X and Y. The more expression patterns or cell types or functional pathways/networks a gene is involved with, the more restrictions would be placed on the number of variants of the gene.

One of the founders of Neo-Darwinism, Ronald Fisher, was one of the first to see that mutations of a given size are more likely to be unfavorable in complex organisms than in simple organisms [40, 41]. The tissue-driven hypothesis has been recently proposed to explain the well-established phenomenon of tissue constraints on mutations [42].

Genetic diversity is restricted by epigenetic complexity and vice versa. It is impossible to build complex epigenetic programs if the DNAs are constantly changing. To compensate for the loss in the range of genetic diversity, complex organisms use different epigenetic programming of the same gene set, in addition to mutation, to adapt to environments and to evolve new phenotypes. Fish and human share nearly identical gene sets and the evolution from fish to human is in a large part a process of epigenetic programming, analogous to writing distinct books with the same set of vocabulary.

Histones are the building blocks for carrying large amount of epigenetic information as posttranslational modifications of histones [43]. In order to have a consistent information coding system, it is intuitively clear that the building blocks for the carrying and writing of the code should be kept constant or unchanging. For example, the four nucleotides for the DNA code have stayed the same throughout evolution. If histones are carriers of information, just like the four nucleotides, their primary sequence should be very stable. Indeed, histones are among the most conserved proteins (H3 is more conserved than EF1 among eukaryotes even though EF1 but not H3 is also found in prokaryotes). This represents an example of the inverse relationship between DNA diversity (DNA coding for histones) and epigenetic complexity. Without epigenetic information conferred upon the histones, the DNA sequence encoding the histones would have changed much more than what has been observed. Likewise, some non-coding DNA sequences are carriers of epigenetic information and should have less freedom to change than other DNA sequences that do not carry such information. The DNAs in complex organisms carry more epigenetic information than DNAs in simple organisms and are therefore less free to change.

Epigenetic restriction of genetic diversity

Research on epigenetic programs is still at its infancy. Based on the limited knowledge of today, we can still envision several ways by which epigenetic programs may restrict genetic diversity. First, most genes are needed for the proper functioning of multiple fetal and adult tissues. A germline mutation in these genes needs to be compatible with multiple tissue types. Thus, the number of viable mutant variants is limited by the number of tissue types with which the gene is involved.

Second, some genes are only expressed in one tissue type, such as hemoglobin in red blood cells. These genes however still exhibit different expression patterns at different time points during development. The gene expression pattern of fetal red blood cells is different from adult red blood cells. So these genes still need to be compatible with several different developmental gene expression patterns. Furthermore, they need to be repressed in most cell types during development and during normal adult life. They need to be packaged into a chromatin state that silences gene expression. Some mutant variants may interfere with such chromatin mediated repression and would be negatively selected.

Third, some genes are expressed in only one cell type but the function of the gene is needed for most cell types of an organism. The function of hemoglobin is needed for the oxygen supply of every cell type. Also, many housekeeping genes such as actin are needed for most cell types. Such general function of a protein like hemoglobin and actin may be fine-tuned for the need of multiple tissues. A housekeeping gene may also exhibit new functions or connections with new networks in complex organisms that are absent in simple organisms, such as the apoptosis function of cytochrome c. Also, for a complex organism to evolve a new cell type, it is necessary to keep the housekeeping genes unchanged so that new cell types can evolve with the least amount of unnecessary disruption to existing cell types. It may not matter much as to which specific version of a housekeeping gene is used but it is important to stick with one once it is selected by an organism.

Fourth, the coding region of every gene in complex organisms encodes not only amino acids but also epigenetic information such as the nucleosome code [44]. A nucleosome code allows the nucleosome to locate in the

right position in the genome. A silent mutation may nevertheless affect the nucleosome code and alters the chromatin packaging state of the gene, which may affect either gene repression or activation. Indeed, nucleosome code has been found to have a negative impact on protein sequence variations [45, 46].

Fifth, complex organisms can eliminate reproductive cells carrying severe mutations [47]. Also, embryos of complex organisms may die or be aborted before birth if they did not develop properly due to mutations.

Sixth, epigenetic enzymes execute a senescence response to oncogenic mutations, thus nullifying the harmful effects of such mutations [48].

Finally, the non-coding and non-expressed regions of the genome are nevertheless packaged into chromatin and encode the nucleosome code and other information necessary for gene expression and organization, and are therefore not free from epigenetic restrictions. Many epigenetic proteins interact with the genome in a sequence specific fashion such as the PRDM family that contains DNA-binding zinc-finger motifs [31]. Even when an epigenetic enzyme has no intrinsic DNA binding property, it nevertheless interacts with a DNA binding transcription factor and therefore requires a specific DNA motif to function as either coactivators or corepressors [49]. Indeed, many non-coding sequences are found under purifying selection [50, 51].

The maximum genetic diversity (MGD) hypothesis

The inverse relationship between genetic diversity and epigenetic complexity is logically and self-evidently true on its own, just like the original intuition that triggered it. It in turn logically deduces what may be termed the maximum genetic diversity (MGD) hypothesis. The hypothesis has three themes. First, for any given organism of certain epigenetic complexity, it can undergo with time either epigenetic changes or genetic mutations within a certain range allowed by the epigenetic complexity. Epigenetic changes often increase epigenetic complexity or organismal complexity, which is termed macroevolution. Genetic changes or mutations cause minor variations in phenotypes and often do not affect the epigenetic programs, which is termed microevolution. Indeed, empirical facts of evolution show both macroevolution and microevolution (Figure 1). The overall direction towards higher complexity however does not necessarily exclude occasionally going in the opposite direction. An organism is more complex if it has a higher degree of epigenetic complexity as indicated by its number of cell types or its number of epigenetic genes. Unlike macroevolution, microevolution is a gradual process of accumulating mutations due to either drift or selection as described by a watered down version of the molecular clock hypothesis or the 'Neutral Theory' and the Neo-Darwinian selection hypothesis. It may also involve some low degree of epigenetic reprogramming without a significant net change in epigenetic complexity.

Second, complex organisms are constructed more by epigenetic programs relative to simple organisms and are in turn inherently less tolerant of mutations. The maximum genetic diversity allowed for a complex organism is smaller than that allowed for a simple organism. Most of the shared residues between two species are due to shared functions and shared epigenetic complexity. A small fraction of the shared residues may be due to common adaptation to a common environmental selection that may vary from time to time (Figure 2). For two distinctly different kinds of organisms over long evolutionary time, their genetic distance is independent of mutation rates and time but is determined by the maximum genetic diversity of the simpler organism of the two. The gradual but stepwise increase in epigenetic complexity with time during macroevolution of distinct organisms results in the near linear correlation between maximum genetic distance and time of species divergence. Such a correlation holds only for macroevolution and is not related to actual mutation rates. It is fundamentally different from the correlation between genetic distance (prior to reaching maximum) and time of divergence during short time scales or before reaching maximum in genetic distance. Actual mutation rates are usually fast enough for maximum genetic distance to be reachable in evolutionary time, especially for fast evolving genes.

Finally, while both micro- and macro-evolution involve gradual accumulation of mutations and minor variations in epigenetic complexity, macroevolution from simple to complex organisms is associated with a punctuational increase in epigenetic complexity and in turn a punctuational loss in genetic diversity (Figure 2 and 3). From a common ancestor, the genetic distance between two splitting descendants may gradually increase with time until reaching a maximum level. This maximum genetic distance will stay roughly unchanged with time thereafter (Figure 3). Mutations still occur but only affect saturated sites or sites that suffer repeated hits. For microevolution, no major changes in epigenetic complexity will take place in either of the two splitting species. For macroevolution, one of the two splitting organisms may undergo a sudden increase in epigenetic complexity. The sudden increase in epigenetic complexity may be a response to the inadequacy of mutation alone in adapting to new environmental challenges. This punctuational jump in epigenetic complexity forces the genetic diversity of the new species to be lower than its sister species that remains largely unchanged in epigenetic complexity. This in turn causes the genetic distance between the new species and its simpler sister species to be strictly determined by the maximum genetic diversity of the sister species over long evolutionary time.

In essence, macroevolution is not at all prolonged microevolution as is assumed by Neo-Darwinism. The two processes are in fact exact opposites with one leading to lesser genetic diversity while the other to greater genetic diversity. What is good for microevolution, i.e., mutation, is mostly bad for macroevolution and must be suppressed in order to evolve higher complexity. Neo-Darwinism is really just a theory of microevolution or population genetics. The MGD hypothesis is a more complete theory that fully grants Neo-Darwinism to be what it really is.

The maximum genetic diversity hypothesis explains numerous facts

The more powerful and fundamental the theory, the more facts it explains. No existing theory of evolution can explain more than half of all facts. In contrast, all the major

facts of evolution can now be easily explained by the MGD hypothesis and a selected few are shown in the following to further illustrate the hypothesis (see Supplementary Information for more facts). In addition, a few novel facts have been uncovered that would represent confirmations of the predictions of the hypothesis. None of these observations are needed to invoke the hypothesis in the first place, since the hypothesis was deduced from intuition or axiom. Therefore, all of them can be considered as independent lines of evidence in support of the hypothesis.

1. *Relationship between genetic diversity and time of origin.* It is well established that genetic diversity within a biological kind of old lineage is greater than that within a biological kind of young lineage (Figure 4A). The genetic diversity of bacteria is greater than eukaryotes [52]. The fact that simple organisms with inherently high-level tolerance of genetic diversity evolved earlier in history generates the apparent correlation between the time of origin and genetic diversity (Figure 4A). But an equally valid relationship is between the time of origin and the epigenetic complexity of the organism (Figure 4B). If epigenetic complexity sets up a maximum cap on genetic diversity and if simple organisms appeared earlier than complex organisms, then the apparent correlation between time of origin and genetic diversity can be explained as an epiphenomenon of epigenetic complexity that is largely independent of mutation rates, generation times, and population size.

2. *The MGD hypothesis predicts the genetic equidistance result.* The equidistance to an outgroup shared by all sister species of a more complex clade is mostly determined by the maximum genetic diversity of the outgroup, which is larger than the maximum genetic diversity of the more complex clade. This notion is illustrated by a hypothetical case as shown in Table 1. If amphibian is allowed a maximum diversity of 60% difference in a hypothetical protein sequence of 10 amino acids as shown in Table 1, then amphibian 1 would differ from a maximum diverged amphibian 2 in 6 of the 10 amino acid positions. Amphibian is the simpler outgroup to the mammalian clade and all mammalian sister species have lower genetic diversity than amphibians. So, the distance between amphibian and mammalian species is mostly determined by the genetic diversity of amphibians, which is maximum 60% difference in this case (Table 1).

It is well known that sequence regions conserved in simple organisms are often also conserved in complex organisms. Sequence regions not conserved in complex organisms are also often not conserved in simple organisms. This explains the fact as illustrated in Table 1 that a comparison of amphibian (with a hypothetical maximum diversity of 60%) and human (with a hypothetical maximum diversity of 10%) should result in a dissimilarity of 60% equaling the maximum diversity of amphibian, rather than 70%. However, it is possible for the variation in humans to contribute a small part to the distance with amphibians. Similarly, the variation in mice may also contribute a small part to the distance with amphibians. Because mice have more genetic diversity than humans, mice would contribute more to the distance with amphibians than humans do. Thus, while mice and humans are approximately equidistant to amphibians, mice may be slightly more distant to amphibians than humans are. But this slight difference may only become apparent when large number of genes or sequences is analyzed.

This notion that the maximum genetic diversity of a simple kind of outgroup organism determines the distance between the outgroup and the more complex clade can be illustrated by the example of cytochrome c. The maximum diversity in this protein sequence is about 70% difference within bacteria, for example, between *Bordetella parapertussis* and *Paracoccus Versutus*. The maximum distance between bacteria and mammals is about 65% difference, such as between *Bordetella parapertussis* and *Pan troglodytes* (chimpanzees). Within fungi, the maximum diversity is about 40% difference, for example, between *Aspergillus oryzae* and *Yarrowia lipolytica*. The maximum distance between fungi and mammals is about 43% difference, such as between *Aspergillus oryzae* and *Pan troglodytes*. Within arthropods, the maximum diversity is about 24% difference, for example, between *Drosophila melanogaster* and *Tigriopus californicus.* The maximum distance between arthropods and mammals is about 25% difference, such as between *Drosophila melanogaster* and *Pan troglodytes*.

This explanation of the genetic equidistance result by the MGD hypothesis can also be easily illustrated by a simple thought experiment. If we can create a yeast, a fish, and a human being by using identical genes for their shared homologs and let the three organisms diverge for an infinite amount of time or about 500 million years, a gene in yeast would have changed a lot to a maximum of, say 50%, while its homolog in fish would have changed to a maximum of, say, 30%, and its homolog in human would have changed very little, say less than 1%. Any more changes than 50% would be lethal to yeast; any more changes than 30% would be lethal to fishes; and any more changes than 1% would be lethal to humans. The reason that a gene in yeast can change much more than in fish, which is still more than in human, is because a gene in human encounters far more functional constraints than its homolog in fish or in yeast. Thus the genetic distance between yeast and human or fish is mainly determined by the mutations in yeast. In this case, the 50% change in yeast would account for the genetic distance of 50% identity between yeast and human or between yeast and fish, as well as 50% identity in within species distance in yeast. The 30% change in fish would account for the genetic distance of 30% identity between fish and human. In contrast, the modern evolution theory would predict that both human and fish can also, like yeast, change up to 50% and would have a genetic distance of 50% identity.

According to the MGD hypothesis, the different residues between yeast and human may be mostly neutral changes for the yeast. But most of these different residues would not be neutral for humans. The mistake of the modern evolution theory is to assume that these different residues between yeast and human are equally neutral changes for both yeast and human. The theory however is relevant only for two diverging species that have similar degree of functional constraints. It is correct to say that the 50% changes between two substrains of yeasts are mostly neutral changes for both substrains. So the modern evolution theory is a theory of microevolution, relevant only to

divergence of similar or identical organisms. It does not describe macroevolution because it fails to take into account the obvious fact that functional constraints on mutations differ tremendously in different kinds of organisms. A theory of microevolution is necessarily unsuitable for macroevolution because the two different evolutionary processes are complete opposites. Neo-Darwinism was from the beginning a theory of population genetics or microevolution invented by geneticists of the 1940s who had essentially no understanding of epigenetics.

If fishes, frogs, birds, and mammals were all created at the same time 450 million years ago by using identical genes for their shared homologs, it would give us the same genetic relationship of these species as we actually observe today. So, the fact that mammals are closer to birds than to frogs in molecular sequence cannot by itself be used to conclude that mammals and birds had a more recent common ancestor than mammals and frogs did. We can only use fossil records for such inference. The phenomenon of maximum genetic diversity invalidates the fundamental notion of the modern evolution theory that sequence similarity can be used to infer time of divergence regardless of length of evolutionary time.

3. *Falsifying the constant clock interpretation of the genetic equidistance result: mammals are closer to birds than to snakes.* If the constant clock interpretation of the genetic equidistance result is true, the complexity of the outgroup should make no difference to the equidistance result. But the MGD hypothesis predicts that the equidistance result only holds when the outgroup is less complex than the sister species. According to the MGD hypothesis, the genetic distance between a complex outgroup and a simple taxon is mainly determined by the genetic diversity of the simple taxon. If one of the sister taxa is more complex than the others, it would have lower genetic diversity and thus smaller genetic distance to a more complex outgroup species.

Mammals and reptiles (including birds) were separated ~310 MyBP. Thus, all the reptiles (including birds) should be equidistant to a mammal if the constant mutation rate idea is true. But the MGD hypothesis predicts that simpler reptiles such as snakes, which lost limbs, should have higher genetic diversity and hence be more distant to a mammal than complex reptiles such as birds. A random sampling of 23 proteins indeed shows that snakes are more distant to humans than birds are in all 23 proteins examined ($P < 0.01$, ND1-ND6, Cox1-Cox3, COB, CytC, HBA, HBB, albumin, ACTB, MC1R, ENO1, FBP1, Mos, Rag1, Rag2, Jun, Adam1a). Snakes are also more distant to humans than lizards are.

This new result, termed genetic non-equidistance to a more complex outgroup, is in contrast to the 45 year old result of genetic equidistance to a simpler outgroup. Only the MGD hypothesis can explain both of these results while the constant clock can only explain one of these. This new result of genetic non-equidistance to a more complex outgroup is extremely robust and universal (Huang, manuscript in preparation), and has important implications for molecular phylogeny studies. It suggests that closer sequence similarity to a complex organism such as humans cannot by itself be used to infer closer genealogy to humans.

One prominent result that immediately comes to mind is the sister relationship between humans and chimpanzees. This relationship is purely based on the fact that chimpanzee is closest to human in sequence similarity than any other animals. However, the rationale for grouping humans and chimpanzees as one clade to the exclusion of other great apes would equally justify the absurd grouping of humans and birds as one clade to the exclusion of snakes. Therefore, the presently popular grouping of humans and chimpanzees is not based on sound rationale and needs to be reevaluated by new analysis that is not solely based on sequence similarity to humans (Huang, manuscript in preparation). The possibility of a great ape clade (containing chimpanzees, gorillas, and orangutans) with human as the outgroup is real and fully compatible with the closer distance between humans and chimpanzees because chimpanzees are more complex or intelligent than other great apes.

4. *Anomalies of the genetic equidistance result*: *frogs are closer to birds than to snakes.* Frog is the simple outgroup relative to the reptile/bird clade. But snake may be the simplest organism within the reptile/bird clade and may even be simpler than the outgroup frog, since it has lost all four limbs. If so, the MGD hypothesis would predict that snakes should be more distant to frogs than birds are, while the molecular clock would predict equidistance. My analysis showed that frogs (*X. laevis*) are closer to birds than to snakes in all 14 randomly selected proteins ($P < 0.01$, Cox1, Cox2, Cox3, ND1, ND2, ND3, ND6, albumin, HBA, HBB, ACTB, ENO1, FBP1, RAG1). Similar analysis also showed that frogs are closer to lizards than to snakes. Furthermore, humans are closer to frogs than to snakes (unpublished), indicating that the genetic diversity of snakes is indeed greater than that of frogs or that snakes are less complex than frogs.

Evolution towards lower complexity is not uncommon. Loss of sight in blind cave fish is another example. Human is closer to zebrafish than to the blind fish *Astyanax mexicanus* in all 11 randomly examined proteins, suggesting higher genetic diversity and lower complexity of blind fish relative to regular fish (Pax6, Ptc2, Dlx3b, Alpha-a-crystallin, Opsin1, Hsp90AA1, Pax2a, NKX2-4, ND2, Fgf8, Oca2). The outgroup must be simpler than all the sister species of a more complex clade in order for the equidistance result to hold. The MGD hypothesis explains easily the violations to the equidistance result while the molecular clock cannot. The typical response in the molecular evolution field to a contradiction like snakes is to speculate that snakes have faster mutation rate. But such ad hoc speculation is a tautology and has no independent evidence. It is also incoherent with the claim of the molecular clock hypothesis that different species should have similar mutation rate. By contrast, the explanation offered by the MGD hypothesis that snakes are less complex is internally coherent and independently supported by the fact that snakes have lost limbs. That limbless represents lower complexity is not only intuitively true but is also supported by the fact that limbless amphibians such as caecilians show greater distance to humans than frogs do (Huang, unpublished).

A small number of genes show anomalies and are routinely excluded from phylogenetic analysis based on the molecular clock hypothesis. An example is the mitochon-

drial protein ND6. My analysis showed that all vertebrates ND6 proteins are equidistant to the outgroup sea urchin but fishes ND6 proteins are closer to frogs than to mammals. The molecular clock hypothesis has no explanation for such a gene that shows both equidistance as well as non-equidistance. However, the MGD hypothesis easily explains it. Some of the shared sequences are due to common environmental selections (Figure 2). Fishes may have more in common with frogs than with mammals in their adaptation strategies for the ND6 protein.

5. *The relationship between time and genetic distance in microevolution is different from that between time and maximum genetic distance in macroevolution.* Most genes (about 90%) have been found to behave consistently as good clocks in macroevolution, and show the same pattern as originally found for cytochrome c [2, 53]: human is more related to primates, less to rodents, still less to birds, still less to frogs, and still less to fish (e.g., see Table 2). However, despite their consistent pattern in macroevolution, many genes give erratic or contradictory results when the timing of split in microevolution is measured. For example, pufferfish (*Takifugu rubripes*) and zebrafish (*Danio rerio*) are believed to have diverged not more than 140-200 MyBP (million years before present) based on the first fossil evidence of teleostei in the early Cretaceous period [54]. If the situation between the two fishes is similar to what one originally found for cytochrome c in macroevolution, one would expect 90% of all genes to show more identity between the fishes than between human and bird since the time of divergence for human and bird is much earlier (310 MyBP).

In a survey of 40 randomly picked genes, I found 36 (90%) that show the expected macroevolution pattern where human is more related to bird, less to frog, and still less to fish. In contrast, only 19 (48%) show more identity between the two fishes than between human and bird. Depending on which gene is used as clock, the time of divergence between the two fishes would vary from 91 to 420 million years (Table 2). In fact, I employed the molecular clock method to derive an average time of divergence using these 40 genes by calibrating against the fossil divergence time between human and bird (310 MyBP). However, I obtained an obviously incorrect time (417+/-172 MyBP) that is more than two fold greater than the actual time as indicated by the fossil record. As a positive control to show that my method is similar to those of others, I derived a mean time of divergence between human and amphibians and found it to be similar to that obtained by others [55].

Apparently, some of the subspecies split or microevolution is not equivalent to the changes in macroevolution, but the Neo-Darwinian hypothesis treats them the same. In contrast, the MGD hypothesis considers them to be very different in evolutionary dynamics. So, clocks derived from macroevolution should not be expected to work also for microevolution. Genetic distance between two distinct species of macroevolution mostly reflects the maximum genetic distance, especially for fast evolving genes. However, genetic distance between two similar species that have diverged more recently would gradually increase as a function of time before it reaches the maximum (Figure 2). Different genes would diverge according to different mutation rates. If the time is not enough for all genes to reach the maximum diversity level, some genes may reach a diversity level closer to the maximum than some other genes. The genes in fish are allowed a maximum diversity level greater than genes in birds and humans. So if some fast evolving genes reached a diversity level closer to the maximum, they would put the time of split between the two fishes earlier than that between birds and humans. But some other slow evolving genes may only reach a certain diversity level much lower than the maximum because of slower rate of mutations and insufficient time. These genes would put the time of split between the fishes later than that between birds and humans.

6. *Simple organisms show higher genetic diversity than complex organisms after evolving for the same amount of time.* The MGD hypothesis predicts that simple organisms should show higher genetic diversity than complex organisms after the same amount of time of evolution. Indeed, flowering plants have much greater genetic diversity than mammals even though they have both coevolved for similar amount of time [4]. Flowering plants are less complex in epigenetic programs and have zero PRDM family of epigenetic enzymes while mammals have 16 to 17. It is also obvious that flowering plants have less number of cell types than mammals. Also, two different mice strains that separated no more than 12 million years ago had more dissimilarity in DNA than human and monkey that shared a common ancestor 20-30 million years ago [56]. At the DNA sequence level, Apodemus and Mus differ by 18% as estimated from neutral sites of genes. In comparison, genome divergence is 8% between human and the Old World monkeys.

7. *Direct evidence of maximum genetic diversity.* The MGD hypothesis predicts that the genetic distance between some ancient species of similar kind or epigenetic complexity may have reached a maximum cap long before present. I tested this prediction for the fungi kingdom. The baker's yeast *Saccharomyces cerevisiae* belongs to the *Ascomycota* phylum, the *Saccharomycotina* subphylum, the *Saccharomycetes* class, the *Saccharomycetales* order, the *Saccharomycetaceae* family, and the *Saccharomyces* genus. A large number of observations have established the well-known top-down direction of evolution where the major pulse of divergence of phyla occurs before subphyla or classes, classes before that of order, orders before that of families, and families before that of genera. If many fungi may share similar epigenetic complexity, the MGD hypothesis predicts that, if time is long enough for genetic distance to reach the cap, the maximum genetic distance between two fungi genera of the same family should be similar to that between two fungi families, or orders, or phyla. In contrast, the molecular clock hypothesis predicts that the genetic distance between two fungi genera of the same family should be smaller than that between families, still smaller than that between orders, still smaller than that between classes or subphyla, and still smaller than that between phyla.

I randomly picked three proteins for analysis, Pin1, CytC, and CMD. As shown in Table 3, the protein sequence identity in Pin1 between two distant genera (*S. cerevisiae* and *D. hansenii)* of the same family is 44%, which is about the same as that between two families of the same order

(39% between *S. cerevisiae* and *Y. lipolytica*), or about the same as that between two subphyla of the same phylum (42% between *S. cerevisiae* and *G. zeae PH-1*), or about the same as that between two phyla of the same kingdom *(41% between S. cerevisiae* and *C. Neo-formans*). For CytC, the identity between two distant genera (78%) seems to be larger than that between two distant families (73%) which is still larger than that between two distant subphyla (67%) which is still larger than that between two distant phyla (60%) (Table 3). This pattern is consistent with the top down direction of evolution and suggests that the time may not yet be long enough for the genetic distance in CytC among the presently sequenced fungi taxa to reach the maximum cap, consistent with the known slow mutation rate of the CytC protein. For the protein CMD, genetic distance between taxa above the family level appears to have reached a maximum at 56-60% identity. These data show that there is a maximum cap on genetic distance at some faster mutating loci like Pin1 and CMD between two species of similar kind in the fungi kingdom. The cap may be gradually reached by gradual accumulation of mutations within a certain amount of time.

I also found direct evidence of maximum cap in fishes. Zebrafish and pufferfish diverged not more than 140-200 MyBP ago as mentioned above. If they diverged by the gradual model and if time is long enough for at least some genes to reach the maximum genetic distance, the MGD hypothesis predicts that some genes would show a genetic distance between the two fishes that is similar to the maximum genetic diversity allowed for fishes. The maximum genetic diversity of fishes is of course roughly the same as the genetic distance between fishes and a distinct fish descendant such as a mammal. I examined a large number of chromatin modifying enzymes and found 13 out of 32 with a distance between the fishes to be the same or slightly greater than the distance between a fish and a mammal (Table 4). The SET family of histone lysine methyltransferases (KMTs) is specifically more enriched with genes that evolved fast with 6 out of 9 genes analyzed reaching maximum cap in the fishes. This feature of the KMT family is significantly different from a slowly evolving family such as ribosomal proteins with only 2 of 12 proteins analyzed reaching maximum cap in the fishes ($P < 0.05$, Fisher's exact test, two tailed). Not a single gene was found to have significantly greater distance between the two fishes than between fish and mammal, indicating clearly the existence of a cap on genetic distance.

8. *Cancer as a disease of both genetics and epigenetics.* The MGD hypothesis predicts that high epigenetic complexity has a way of limiting the incidence of mutations. A relaxation in epigenetic control may be expected to allow more mutations to occur. Indeed, human cancer provides a good illustration of this prediction [57]. Mutations are common in cancer. Epigenetic programs are often deregulated in cancer and methylation deficiency is a hallmark of cancer [31, 57, 58]. Loss of epigenetic control as indicated by loss of DNA methylation occurs during aging and precedes mutations in cancer [59]. A rate-limiting step in carcinogenesis by major environmental factors such as nutrient-imbalanced diet is the deregulation of an epigenetic enzyme RIZ1/PRDM2 [60]. In addition, the hypothesis predicts that high genetic diversity or too many mutations would interfere with epigenetic programming. Indeed, too many mutations, either germ line or somatic, are well known to cause cancer, which is essentially a disease where the normal epigenetic programs have been replaced by a cancer specific program. Thus, the hypothesis unifies cancer genetics and epigenetics and explains why cancer appears to be a disease of both genetic mutations and epigenetic anomalies.

The hypothesis explains species difference in cancer rate. The low epigenetic complexity state of simple organisms can tolerate more mutations or should be less prone to transformations caused by random mutations. Indeed, cancer is rare in the long-lived turtles or in non-human primates. Since spontaneous random mutations occur more or less similarly in different organisms, the rate-limiting factor for the difference in cancer incidence cannot possibly be mutations and must be the complexity level of epigenetic programs. Also, complex organisms have more ways of restricting mutations and yet still have higher cancer rate than simple organisms. This again suggests that the rate-limiting factor here is not mutations. Finally, more tissues are prone to cancer in humans than in simple organisms such as rodents, consistent with the fact that more cell types in complex organisms are epigenetically complex and thus more prone to cancer [61].

9. *Genetic diseases.* The prevalence of genetic or familial diseases in humans indicates plainly that a large portion of genetic diversity, i.e. those represented by those disease mutations, cannot become a part of the normal range of genetic diversity among humans. Most genetic diseases affect only a tiny population of humans. Just imagine how much more diversified the human race would be if all those rare disease mutations would become fixed in the whole population. If mutations in the retinoblastoma gene, a cell cycle regulator important for many different cell types, do not cause cancer in the retina of children, the diversity in the retinoblastoma gene locus would be greatly expanded. The fact of rare disease mutations in humans is sufficient to prove the hypothesis that there is an upper limit to the amount of genetic diversity in an organism. The fact that those rare disease mutations are mostly tissue specific is consistent with the notion that the upper limit is set up by the complexity of epigenetic programs. If humans lack the retina cell type or the retina specific epigenetic program, most of the mutations in the retinoblastoma gene would have been tolerated as normal variations and the genetic diversity of humans would have been in turn expanded. Also, numerous disease alleles in humans correspond to normal alleles in rhesus macaques [62]. Thus, many alleles or mutant variants that can be tolerated in a less complex organism in fact cause diseases in humans.

10. *Ubiquitously expressed genes have lower genetic diversity than tissue specific genes.* The MGD hypothesis predicts that ubiquitously expressed genes have lower diversity than tissue specific genes due to selection against mutations that cannot fit with multiple cell types. Indeed, an analysis of 2400 genes between human and rodent found that ubiquitously expressed proteins have average genetic distances between human and rodent that were threefold lower than those of tissue-specific genes [63]. The effects of tissue constraints on sequence variation have been independently found by others [42, 64, 65]. Furthermore,

housekeeping genes tend to cluster in low mutation regions of a chromosome [66].

Brain obviously contains many more cell types than other organs such as testis or liver. Thus, brain specific genes are expressed in more cell types than testis specific genes and therefore less likely to change. Indeed, brain specific genes evolve slower than other tissue specific genes such as testis and liver [42, 67-75]. That brain specific genes behave like housekeeping genes in evolution rate supports the notion that brain has many distinct cell types. Also, genes highly expressed in the prefrontal cortex evolve the slowest among all of 78 tissues examined (including 21 brain sub regions), correlating well with the known function of this region in complex goal-orientated cognitive tasks [75].

A housekeeping gene in complex organisms is expressed in more tissues than in simple organisms. A brain specific gene in a complex brain is expressed in more neuronal cell types than in a simple brain. The fact that housekeeping genes and brain specific genes evolve slowly proves the main theme of the MGD hypothesis that complex organisms with more cell types and complex brains should have lower genetic diversity.

11. *Evolution towards higher complexity is accompanied by a reduction in reproductive efficiency that is more hostile to mutant embryos.* An efficient reproductive system should have high rate of fertility, low rate of spontaneous abortion, and resistance to aging. An embryo of complex organisms is more sensitive to spontaneous mutations and may be aborted due to mutation-induced abnormal development. Also, older parents accumulate more mutations and are far more likely to give rise to progenies with more mutations. Thus, the low genetic diversity of complex organisms could be in part maintained by a less efficient reproductive system that has low fertility rate, high abortion rate, and high sensitivity to aging.

Indeed, the infertility rate of humans (15% of couples are infertile) is significantly lower than other less complex animals such as mice. In breeding experiments done in my lab, none of 41 couples of mice were infertile, which is significantly lower than 15% ($P = 0.026$, Fisher's exact test, two tailed). Unlike other animals, the spontaneous abortion rate of humans increases dramatically with age. While human and chimpanzees have similar abortion risk of 10% at young age, the abortion risk of older humans is 75%, much higher than chimpanzees of similar age (23%) [76, 77]. Many spontaneous abortions are due to genetic abnormalities of the embryos. Mutations in many genes cause embryonic lethality as shown by genetic experiments in model mammals. Human is also unique in having menopause among primates [78].

The fact that humans have a less efficient reproductive system relative to most other less complex animals ensures low birth rate of mutant humans, thus contributing to the low genetic diversity of humans. While this may negatively impact the capacity of humans to adapt to environments by way of mutations, it may be of no real consequence since we humans primarily use the creativity of our mind to adapt. This phenomenon of low birth rate of mutants is precisely what would be predicted by the MGD hypothesis. And it is precisely what most humans would have wanted since no one wants to have children with birth defects. Most humans would be grateful to mother nature for the great gift of abortion.

In contrast, Darwinism would have predicted abortion to be bad rather than good. Indeed, experts such as Ayala had in fact claimed that the high rate of spontaneous abortions of humans is an imperfect outcome of Darwinian evolution. If that results from divinely inspired anatomy, Ayala said, "God is the greatest abortionist of them all." [79]. Here, Ayala failed to see the virtues of abortion, which is not unexpected since his Darwinian mindset would prefer that all mutants get born. Darwinian natural selection or elimination of mutants is supposed to take place at the age of reproduction of these mutants rather than at pre-birth. Darwin himself predicted that imperfect transitional forms should be abundant during evolution. Indeed, Darwinism would have predicted a significant portion of human populations to be deformed or diseased unfit individuals. The fact that imperfect transitional forms did not exist in the past in the fossils nor in today's world has exactly the same reason. The imperfect forms never got born in any significant number due to a quality control mechanism that is anti-mutation and anti-Darwinism, i.e., abortion.

12. *Human has the lowest genetic diversity.* The genetic diversity within chimpanzees is two to three times greater than within humans, even though both species are thought to have evolved for the same amount of time since their most recent common ancestor [80]. The striking fact that human shows the lowest DNA diversity among all species has commonly been explained by the bottleneck hypothesis: most human populations are thought to go extinct at one time in history except one small population that survived to produce the six billion people living today. But this hypothesis is merely an ad hoc tautology. There is no independent evidence of such near extinction event and there is little hope of ever uncovering such evidence. There are also lines of evidence against the bottleneck hypothesis [81, 82]. In contrast, the homogeneity of human DNA is a precise prediction by the MGD hypothesis. The organism with the most complex and diverse epigenetic programs or the most number of cell types is necessarily supposed to have the lowest diversity in DNA. With everything else being equal, the less efficient reproductive system of humans relative to chimpanzees is sufficient to account for the lower genetic diversity of humans.

Neanderthals appeared earlier than modern humans and have slightly larger brains. It is unclear whether Neanderthals may be less intelligent or have less complex brain than modern humans, which may explain their mysterious extinction. The MGD hypothesis predicts that Neanderthals are less complex in epigenetic programs and have a less complex brain than modern humans because Neanderthals appear to exhibit more DNA diversity [83, 84]. This prediction is obviously consistent with the fact that it is modern humans rather than the Neanderthals that dominate the Earth today. It is also consistent with the evolution trend that less complex organisms appeared earlier in history.

13. *Evidence from fossil DNA and protein sequences.* Finally, the MGD hypothesis explains the recent results that ancient fossil specimens are more distant to an outgroup than extant sister species are in non-neutral sequences [17, 18], and that ancient fossil specimens have greater genetic distance than extant sister species [17]. The Neo-Darwinian

gradual mutation hypothesis predicts that ancient specimens of extinct species cannot be more distant to an outgroup than extant sister species are (Figure 5A). Also, two distinct ancient specimens from different era cannot be more distant than their extant sister species are. But the MGD hypothesis predicts the exact opposite (Figure 5B). The recent analysis of fossil DNA and protein sequences fully conforms to the predictions of the MGD hypothesis rather than to those of the modern evolution theory.

While few extant organisms show violations of the equidistance result, all three fossils (Neanderthals, dinosaurs, and mastodons) for which there are sequence data violate the equidistance result. To explain this 'striking observation', Green et al. invoked the speculative ad hoc idea of less purifying selection due to an imagined small population size for the Neanderthals than modern humans [18]. But this idea has a major contradiction within the framework of the modern evolution theory to which Green et al. subscribe. If the equidistance result is an outcome of constant clock, then it cannot be related to natural selection. The constant clock idea requires that most mutations are neutral. So, if the equidistance result is a consequence of neutral mutations, it simply cannot be a consequence of purifying selections. Therefore, any violations of the equidistance result simply cannot be explained by any selection schemes if one accepts the molecular clock hypothesis.

Implications for molecular phylogeny

Molecular phylogeny analysis aims to classify the time of divergence between morphologically similar species that either do not have fossil records or cannot be clearly distinguished by fossil records. A mutation rate is usually calibrated using fossil records of vertebrate macroevolution. The most commonly used calibration date is the divergence time of 310 million years between birds and mammals. However, as discussed here, the 'mutation rate' deduced from macroevolution is not the real mutation rate as it is known in real time measurements. It therefore cannot be used to time microevolution of species that have diverged only recently and have not yet reached a maximum cap in genetic diversity. Such microevolution reflects the real mutation rate and should be timed using a mutation rate that is measured by real time analysis such as pedigree analysis. For example, to date the divergence of pufferfish and zebrafish, one should only use genes that have not reached a maximum diversity. Thus, cytochrome c may be used but most KMTs should not be used. However, the mutation rate of cytochrome c should be deduced not from divergence of birds and mammals but from pedigree analysis of living pufferfish and zebrafish.

All molecular dating methods rely on the assumption that "If two species have a relatively recent common ancestor, their DNA sequences will be more similar than the DNA sequences for two species that share a distant common ancestor.", as stated by most text books. While this assumption has some factual support, it also has numerous factual exceptions. One important consequence of these exceptions is that they make it impossible to trust the molecular phylogenies constructed by the present methods of molecular analysis. Given the frequent violations of the molecular clock, we cannot know whether any given species may or may not have a constant clock for any given time period. The relative rate test of rate constancy is invalid because it is based on the constant clock interpretation of the genetic equidistance result, which has now been proven false.

Given the sequence dissimilarity in certain genes between two species, one must ask first whether this dissimilarity represents the maximum or not and whether it is contributed mostly by one of the two diverging species. Whenever possible, we should only use slow evolving genes to calculate divergence time since such genes are less likely to have reached maximum distance. It is expected that future studies will show a significant difference between phylogeny derived from slow evolving genes and phylogeny from fast evolving genes. Most studies in the past used average distance from multiple genes. Such studies mostly reflect the results of fast evolving genes since such genes tend to have greater distances and thus contribute more heavily to the calculation of the average distance. It is expected that most major conflicts between molecular dating and fossil dating would disappear when slow evolving genes are used for molecular dating analysis. Since the credibility of the molecular dating method was originally established by its consistency with the fossil record [1, 2, 85], it is self-defeating to distrust the fossil record whenever there is a conflict between fossil dating and molecular dating. From the perspective of the MGD hypothesis, such conflicts are largely due to the present molecular methods being not completely correct.

Molecular clock methods that work for microevolution or divergence of two species with similar epigenetic complexity are not appropriate for macroevolution. Unfortunately, all present molecular methods are for microevolution only and assume that both diverging species contribute equally to the genetic distance between them. Such assumption may be true for microevolution but is clearly false for macroevolution where the genetic distance between two diverging species of different epigenetic complexity is primarily contributed by the simpler species of the two. Any parsimony-based methods of microevolution simply cannot apply to cases where the distance concerned is mostly contributed by one of the two diverging species or represents the maximum.

The genetic equidistance result and molecular clock hypothesis were originally established by using percent identity to measure genetic distance [1, 2, 85]. In most cases of macroevolution or for fast evolving genes, the distance in percent identity represents the maximum. The relationship between such distance and the number of actual mutational events is practically impossible to discern. All known mathematical techniques of converting percent identity into the actual number of mutational events, such as the Poisson Correction or the Gamma distance, have assumptions that may be true only for microevolution prior to reaching maximum distance but are clearly false for macroevolution.

To test for genetic equidistance, the most straight forward and common sense method without any false or uncertain assumptions is to examine the percent identity of a small number of randomly selected genes. Genetic equidistance of A and B to an outgroup C can be established if the number of genes showing greater similarity between A and C than between B and C is similar

to the number of genes showing less similarity between A and C than between B and C ($P > 0.05$). Human is closer to mice than to snakes in almost 100% of all homologous genes and a random sampling of 10 genes is sufficient to establish that mice and snakes are non-equidistant to humans ($P < 0.05$). However, the number of genes showing more similarity between humans and snakes than between mice and snakes is similar to that showing less similarity, and a random sampling of a few dozen genes is sufficient to establish that humans and mice are equidistant to snakes ($P > 0.05$).

The genetic equidistance result has now a sound explanation in the MGD hypothesis and can therefore no longer be used to support the constant clock hypothesis. The molecular clock hypothesis has more evidence against it than for it. The common violation of the molecular clock effectively nullifies its predictive value. We cannot know if a constant clock should or should not apply to any specific groups of species for which we want to establish genetic relationships. Any conclusion that is based on the assumption of similar mutation rates among different species is simply baseless and must be reevaluated by using new methods that do not rely on false or uncertain assumptions.

The most prominent case in need of a reevaluation is the 5-7 million year divergence time between humans and chimpanzees as estimated from molecular clocks [86], which is in sharp conflict with the fossil estimation of ~18 million years [87-92]. The molecular calculation used the mutation rate of monkeys, based on a calibration using fossil divergence time of 25-30 million years between monkey and human, to yield the divergence time of apes and humans. The big assumption here is that monkeys, apes, and humans all have similar mutation rates, which simply has no factual support and is almost certainly false. The molecular clock method is also invalid here because the distance between humans and chimpanzees may be mostly contributed by chimpanzees and may represent the maximum.

The phenomenon of maximum cap on genetic distance means that in most cases we cannot discern the age of a population by using the coalescence method. Based on mutation rates derived from pedigree analysis of the mitochondrial D-loop region, the human race is estimated to be only ~ 6500 years old [93]. However, what the result means is uncertain since the maximum genetic diversity of the mitochondrial D-loop region is unknown for humans. If the genetic diversity has not yet reached a cap within 6500 years, then we may conclude that the human race is indeed 6500 years old. On the other hand, if the cap has been reached in 6500 years, then we will not be able to discern the real age of human race using the mitochondrial D-loop DNA. The age could be much older while the diversity of the D-loop DNA would no longer increase with time after reaching the cap. Given that the oldest fossil of modern humans is much older than 6500 years (about 30,000 to 65,000 years old), it is likely that the maximum diversity of the mitochondrial D-loop can be reached in ~ 6500 years. Thus, if the fossil record is true, the maximum distance in the D-loop that we observe today within the human race would in fact represent the maximum allowable for the human organism.

Conclusions

The inverse relationship between genetic diversity and epigenetic complexity is the first axiom in biology. It does not need independent validation of empirical facts, just like the intuition that a complex system is more selective in building materials than a simple system. Nonetheless, it or its necessary logical deduction, the MGD hypothesis, has found support in numerous facts and has yet to meet a factual contradiction. It explains more facts than does the molecular clock hypothesis or the Neo-Darwinian theory. Many data that were simply ignored before or explained away by ad hoc speculations can now be understood in a coherent way by using a single universal theme. This axiom and the existing theories of macroevolution are mutually exclusive. If one accepts the axiom, which one must as there is no reason not to, then everything else follows as its deductions, including the MGD hypothesis and all of the major facts of evolution.

Most of the existing literature in molecular evolution would need to be reinterpreted in light of the new axiom. The existing theories cannot account for macroevolution and are only relevant to some microevolution events over short timescales. The inference of divergence time based on sequence identity is still practically useful in many cases. But a distinction must be made between divergence that has reached a maximum and divergence that has not. Microevolution must be distinguished from macroevolution. While the genetic distance may be equally contributed by both diverging species during microevolution, it is mostly contributed by the less complex of the two during macroevolution.

Acknowledgments

This work was supported by a grant from the NIH (RO1 CA 105347). I thank Drs. Phil Skell and Klara Briknarova for critical reading of the manuscript.

Notes added in proof

The MGD hypothesis is being confirmed on a constant basis by new findings. The first preprint version of the manuscript was made public on the Internet on April 2, 2008. Two predictions made in that version have now been confirmed by new papers or new knowledge that have since become known to the author. The first is about the restriction of genetic variation by the nucleosome code, which has now been shown by two recent papers [45, 46]. The second is about natural death of mutant embryos prior to birth as a mechanism of epigenetic restriction of mutations. The author was not aware of the fact that human has high rate of spontaneous abortion until reading Ayala's comment on abortion published by the New York Times on April 29, 2008 [79].

Methods

Protein sequences from a specific taxon were retrieved from the NCBI protein database. The exact nature of the genes (function type, reason for study, and time or order of appearance in the Genbank) is independent of the equidistance result. Thus, while the availability of a gene sequence in the Genbank has specific reasons and hence is not strictly random, none of the reasons is in anyway linked to the equidistance result. Their availability in the Genbank

is therefore effectively random as far as the equidistance result is concerned.

Homology comparisons were performed using BLASTP on the NCBI server. Percent nonidentity in protein sequence was used to measure genetic distance as originally used in the 1960s when the genetic equidistance result was first discovered. The equidistance result would not be affected in any way when percent nonidentity was converted into Poisson or Gamma distance. But such conversion is meaningless when the percent nonidentity in fact represents the maximum in distance that has long been reached before present time.

References

1. Zuckerkandl E, Pauling L: **Molecular disease, evolution, and genetic heterogeneity, Horizons in Biochemistry**. New York: Academic Press; 1962.

2. Margoliash E: **Primary structure and evolution of cytochrome c**. *Proc Natl Acad Sci* 1963, **50**:672-679.

3. Kumar S: **Molecular clocks: four decades of evolution**. *Nat Rev Genet* 2005, **6**(8):654-662.

4. Huang S: **The genetic equidistance result of molecular evolution is independent of mutation rates**. *J Comp Sci Syst Biol* 2008, **1**:092-102.

5. Nei M, Kumar S: **Molecular evolution and phylogenetics**. New York: Oxford University Press; 2000.

6. Li W-H: **Molecular evolution**. Sunderland, MA: Sinauer Associates; 1997.

7. Avise JC: **Molecular markers, natural history and evolution**. New York, NY: Springer; 1994.

8. Ayala FJ: **Molecular clock mirages**. *BioEssays* 1999, **21**(1):71-75.

9. Pulquerio MJ, Nichols RA: **Dates from the molecular clock: how wrong can we be?** *Trends Ecol Evol* 2007, **22**(4):180-184.

10. Ho SYW, Larson G: **Molecular clocks: when times are a-changin'**. *Trends Genet* 2006, **22**:79-83.

11. Laird CD, McConaughy BL, McCarthy BJ: **Rate of fixation of nucleotide substitutions in evolution**. *Nature* 1969, **224**(5215):149-154.

12. Jukes TH, Holmquist R: **Evolutionary clock: nonconstancy of rate in different species**. *Science* 1972, **177**(48):530-532.

13. Goodman M, Moore GW, Barnabas J, Matsuda G: **The phylogeny of human globin genes investigated by the maximum parsimony method**. *J Mol Evol* 1974, **3**(1):1-48.

14. Langley CH, Fitch WM: **An examination of the constancy of the rate of molecular evolution**. *J Mol Evol* 1974, **3**(3):161-177.

15. Kasahara M, Naruse K, Sasaki S, Nakatani Y, Qu W, Ahsan B, Yamada T, Nagayasu Y, Doi K, Kasai Y *et al*: **The medaka draft genome and insights into vertebrate genome evolution**. *Nature* 2007, **447**(7145):714-719.

16. Millar CD, Dodd A, Anderson J, Gibb GC, Ritchie PA, Baroni C, Woodhams MD, Hendy MD, Lambert DM: **Mutation and evolutionary rates in adelie penguins from the antarctic**. *PLoS Genet* 2008, **4**(10):e1000209.

17. Huang S: **Ancient fossil specimens are genetically more distant to an outgroup than extant sister species are**. *Riv Biol* 2008, **101**:93-108.

18. Green RE, Malaspinas AS, Krause J, Briggs AW, Johnson PL, Uhler C, Meyer M, Good JM, Maricic T, Stenzel U *et al*: **A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing**. *Cell* 2008, **134**(3):416-426.

19. Yang J, Lusk R, Li WH: **Organismal complexity, protein complexity, and gene duplicability**. *Proc Natl Acad Sci U S A* 2003, **100**(26):15661-15665.

20. Levine M, Tjian R: **Transcription regulation and animal diversity**. *Nature* 2003, **424**(6945):147-151.

21. Vinogradov AE, Anatskaya OV: **Organismal complexity, cell differentiation and gene expression: human over mouse**. *Nucleic Acids Res* 2007, **35**(19):6350-6356.

22. Bonner JT: **Perspective: the size-complexity rule**. *Evolution* 2004, **58**:1883-1890.

23. Carroll SB: **Chance and necessity: the evolution of morphological complexity and diversity**. *Nature* 2001, **409**(6823):1102-1109.

24. McShea DW: **Metazoan complexity and evolution: is there a trend?** *Evolution* 1996, **50**:477-492.

25. Bonner JT: **The evolution of complexity**. Princeton, NJ: Princeton University Press; 1988.

26. Vogel C, Chothia C: **Protein family expansions and biological complexity**. *PLoS Comput Biol* 2006, **2**(5):e48.

27. Anway MD, Cupp AS, Uzumcu M, Skinner MK: **Epigenetic transgenerational actions of endocrine disruptors and male fertility**. *Science* 2005, **308**(5727):1466-1469.

28. Hitchins MP, Wong JJ, Suthers G, Suter CM, Martin DI, Hawkins NJ, Ward RL: **Inheritance of a cancer-associated MLH1 germ-line epimutation**. *N Engl J Med* 2007, **356**(7):697-705.

29. Cropley JE, Suter CM, Beckman KB, Martin DI: **Germ-line epigenetic modification of the murine A vy allele by nutritional supplementation**. *Proc Natl Acad Sci U S A* 2006, **103**(46):17308-17312.

30. Fumasoni I, Meani N, Rambaldi D, Scafetta G, Alcalay M, Ciccarelli FD: **Family expansion and gene rearrangements contributed to the functional specialization of PRDM genes in vertebrates**. *BMC Evol Biol* 2007, **7**:187.

31. Huang S: **Histone methyltransferases, diet nutrients, and tumor suppressors**. *Nat Rev Cancer* 2002, **2**:469-476.

32. Wilson GM, Flibotte S, Missirlis PI, Marra MA, Jones S, Thornton K, Clark AG, Holt RA: **Identification by full-coverage array CGH of human DNA copy number increases relative to chimpanzee and gorilla**. *Genome Res* 2006, **16**(2):173-181.

33. Heimberg AM, Sempere LF, Moy VN, Donoghue PC, Peterson KJ: **MicroRNAs and the advent of vertebrate morphological complexity**. *Proc Natl Acad Sci U S A* 2008, **105**(8):2946-2950.

34. Grimson A, Srivastava M, Fahey B, Woodcroft BJ, Chiang HR, King N, Degnan BM, Rokhsar DS, Bartel DP: **Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals**. *Nature* 2008, **455**(7217):1193-1197.

35. Taft RJ, Pheasant M, Mattick JS: **The relationship between non-protein-coding DNA and eukaryotic complexity**. *Bioessays* 2007, **29**(3):288-299.

36. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB: **Alternative isoform regulation in human tissue transcriptomes**. *Nature* 2008, **456**:470-476.

37. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ: **Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing**. *Nat Genet* 2008, **40**:1413-1415.

38. Stevens CF: **Neuronal diversity: too many cell types for comfort?** *Curr Biol* 1998, **8**(20):R708-710.

39. Enard W, Fassbender A, Model F, Adorjan P, Paabo S, Olek A: **Differences in DNA methylation patterns between humans and chimpanzees**. *Curr Biol* 2004, **14**(4):R148-149.

40. Fisher RA: **The genetical theory of natural selection.** Oxford, U.K.: Oxford University Press; 1930.

41. Orr HA: **Adaptation and the cost of complexity**. *Evolution* 2000, **54**(1):13-20.

42. Gu X, Su Z: **Tissue-driven hypothesis of genomic evolution and sequence-expression correlations**. *Proc Natl Acad Sci U S A* 2007, **104**(8):2779-2784.

43. Strahl BD, Allis CD: **The language of covalent histone modifications**. *Nature* 2000, **403**(6765):41-45.

44. Segal E, Fondufe-Mittendorf Y, Chen L, Thastrom A, Field Y, Moore IK, Wang JP, Widom J: **A genomic code for nucleosome positioning**. *Nature* 2006, **442**(7104):772-778.

45. Warnecke T, Batada NN, Hurst LD: **The impact of the nucleosome code on protein-coding sequence evolution in yeast**. *PLoS Genet* 2008, **4**(11):e1000250.

46. Washietl S, Machne R, Goldman N: **Evolutionary footprints of nucleosome positions in yeast**. *Trends Genet* 2008, **24**(12):583-587.

47. Fan W, Waymire KG, Narula N, Li P, Rocher C, Coskun PE, Vannan MA, Narula J, Macgregor GR, Wallace DC: **A mouse model of mitochondrial disease reveals germline selection against severe mtDNA mutations**. *Science* 2008, **319**(5865):958-962.

48. Braig M, Lee S, Loddenkemper C, Rudolph C, Peters AH, Schlegelberger B, Stein H, Dorken B, Jenuwein T, Schmitt CA: **Oncogene-induced senescence as an initial barrier in lymphoma development**. *Nature* 2005, **436**(7051):660-665.

49. Rosenfeld MG, Lunyak VV, Glass CK: **Sensors and signals: a coactivator/corepressor/epigenetic code for integrating signal-dependent programs of transcriptional response**. *Genes Dev* 2006, **20**(11):1405-1428.

50. Andolfatto P: **Adaptive evolution of non-coding DNA in Drosophila**. *Nature* 2005, **437**(7062):1149-1152.

51. McLean C, Bejerano G: **Dispensability of Mammalian DNA**. *Genome Res* 2008, **18**(11):1743-1751.

52. Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P: **Toward automatic reconstruction of a highly resolved tree of life**. *Science* 2006, **311**(5765):1283-1287.

53. Fitch WM, Margoliash E: **Construction of phylogenetic trees**. *Science* 1967, **155**(760):279-284.

54. Powers DA: **Evolutionary genetics of fish**. *Advances in Genetics* 1991, **29**:119-228.

55. Kumar S, Hedges SB: **A molecular timescale for vertebrate evolution**. *Nature* 1998, **392**(6679):917-920.

56. Xiang AP, Mao FF, Li WQ, Park D, Ma BF, Wang T, Vallender TW, Vallender EJ, Zhang L, Lee J *et al*: **Extensive contribution of embryonic stem cells to the development of an evolutionarily divergent host**. *Hum Mol Genet* 2008, **17**(1):27-37.

57. Huang S: **Histone methylation and the initiation of cancer, Cancer Epigenetics**. New York: CRC Press; 2008.

58. Feinberg AP, Tycko B: **The history of cancer epigenetics**. *Nat Rev Cancer* 2004, **4**(2):143-153.

59. Suzuki K, Suzuki I, Leodolter A, Alonso S, Horiuchi S, Yamashita K, Perucho M: **Global DNA demethylation in gastrointestinal cancer is age dependent and precedes genomic damage**. *Cancer Cell* 2006, **9**(3):199-207.

60. Zhou W, Alonso S, Takai D, Lu SC, Yamamoto F, Perucho M, Huang S: **Requirement of RIZ1 for Cancer Prevention by Methyl-Balanced Diet**. *PLoS ONE* 2008, **3**(10):e3390.

61. Anisimov VN, Ukraintseva SV, Yashin AI: **Cancer in rodents: does it tell us about cancer in humans?** *Nat Rev Cancer* 2005, **5**(10):807-819.

62. Gibbs RA, Rogers J, Katze MG, Bumgarner R, Weinstock GM, Mardis ER, Remington KA, Strausberg RL, Venter JC, Wilson RK *et al*: **Evolutionary and biomedical insights from the rhesus macaque genome**. *Science* 2007, **316**(5822):222-234.

63. Duret L, Mouchiroud D: **Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate**. *Mol Biol Evol* 2000, **17**(1):68-74.

64. Zhang L, Li WH: **Mammalian housekeeping genes evolve more slowly than tissue-specific genes**. *Mol Biol Evol* 2004, **21**(2):236-239.

65. Zhu J, He F, Hu S, Yu J: **On the nature of human housekeeping genes**. *Trends Genet* 2008, **24**(10):481-484.

66. Lercher MJ, Urrutia AO, Hurst LD: **Clustering of housekeeping genes provides a unified model of gene order in the human genome**. *Nat Genet* 2002, **31**(2):180-183.

67. Miyata T, Kuma K, Iwabe N, Nikoh N: **A possible link between molecular evolution and tissue evolution demonstrated by tissue specific genes**. *Jpn J Genet* 1994, **69**(5):473-480.

68. Kuma K, Iwabe N, Miyata T: **Functional constraints against variations on molecules from the tissue level: slowly evolving brain-specific genes demonstrated by protein kinase and immunoglobulin supergene families**. *Mol Biol Evol* 1995, **12**(1):123-130.

69. Khaitovich P, Hellmann I, Enard W, Nowick K, Leinweber M, Franz H, Weiss G, Lachmann M, Paabo S: **Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees**. *Science* 2005, **309**(5742):1850-1854.

70. Shi P, Bakewell MA, Zhang J: **Did brain-specific genes evolve faster in humans than in chimpanzees?** *Trends Genet* 2006, **22**(11):608-613.

71. Bakewell MA, Shi P, Zhang J: **More genes underwent positive selection in chimpanzee evolution**

than in human evolution. *Proc Natl Acad Sci U S A* 2007, **104**(18):7489-7494.

72. Kosiol C, Vinar T, da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, Siepel A: **Patterns of positive selection in six Mammalian genomes**. *PLoS Genet* 2008, **4**(8):e1000144.

73. Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, Hubisz MJ, Fledel-Alon A, Tanenbaum DM, Civello D, White TJ *et al*: **A scan for positively selected genes in the genomes of humans and chimpanzees**. *PLoS Biol* 2005, **3**(6):e170.

74. Consortium. CSaA: **Initial sequence of the chimpanzee genome and comparison with the human genome**. *Nature* 2005, **437**(7055):69-87.

75. Tuller T, Kupiec M, Ruppin E: **Evolutionary rate and gene expression across different brain regions**. *Genome Biol* 2008, **9**(9):R142.

76. Nybo Andersen AM, Wohlfahrt J, Christens P, Olsen J, Melbye M: **Maternal age and fetal loss: population based register linkage study**. *BMJ* 2000, **320**(7251):1708-1712.

77. Roof KA, Hopkins WD, Izard MK, Hook M, Schapiro SJ: **Maternal age, parity, and reproductive outcome in captive chimpanzees (Pan troglodytes)**. *Am J Primatol* 2005, **67**(2):199-207.

78. Emery Thompson M, Jones JH, Pusey AE, Brewer-Marsden S, Goodall J, Marsden D, Matsuzawa T, Nishida T, Reynolds V, Sugiyama Y *et al*: **Aging and fertility patterns in wild chimpanzees provide insights into the evolution of menopause**. *Curr Biol* 2007, **17**(24):2150-2156.

79. Dean C: **Roving defender of evolution, and rood for God**. In: *New York Times.* New York: Arthur Ochs Sulzberger, Jr.; 2008: F4.

80. Becquet C, Patterson N, Stone AC, Przeworski M, Reich D: **Genetic structure of chimpanzee populations**. *PLoS Genet* 2007, **3**(4):e66.

81. Li WH, Sadler LA: **Low nucleotide diversity in man**. *Genetics* 1991, **129**(2):513-523.

82. Xiong WJ, Li WH, Posner I, Yamamura T, Yamamoto A, Gotto AM, Jr., Chan L: **No severe bottleneck during human evolution: evidence from two apolipoprotein C-II deficiency alleles**. *Am J Hum Genet* 1991, **48**(2):383-389.

83. Krings M, Capelli C, Tschentscher F, Geisert H, Meyer S, von Haeseler A, Grossschmidt K, Possnert G, Paunovic M, Paabo S: **A view of Neandertal genetic diversity**. *Nat Genet* 2000, **26**(2):144-146.

84. Orlando L, Darlu P, Toussaint M, Bonjean D, Otte M, Hanni C: **Revisiting Neandertal diversity with a 100,000 year old mtDNA sequence**. *Curr Biol* 2006, **16**(11):R400-402.

85. Doolittle RF, Blombaeck B: **Amino-Acid Sequence Investigations Of Fibrinopeptides From Various Mammals: Evolutionary Implications**. *Nature* 1964, **202**:147-152.

86. Wilson AC, Sarich VM: **A molecular time scale for human evolution**. *Proc Natl Acad Sci U S A* 1969, **63**(4):1088-1093.

87. Simons EL: **The phyletic position of Ramapithecus.** . *Postilla* 1961, **57**:1-9.

88. Simons EL, Pilbeam DR: **Preliminary revision of the Dryopithecinae (Pongidae, Anthropoidea)**. *Folia Primatol (Basel)* 1965, **3**(2):81-152.

89. Pilbeam D: **The earliest hominids**. *Nature* 1968, **219**(5161):1335-1338.

90. Schwartz JH: **The evolutionary relationships of man and orang-utans**. *Nature* 1984, **308**(5959):501-505.

91. Lewin R: **Human Evolution**, 5 th edn. Malden, MA 02148, USA: Blackwell Publishing Ltd; 2005.

92. Schwartz JH: **The Red Ape, Orangutans and Human Origins**. Cambridge, MA 02142, USA: Westview Press; 2005.

93. Parsons TJ, Muniec DS, Sullivan K, Woodyatt N, Alliston-Greiner R, Wilson MR, Berry DL, Holland KA, Weedn VW, Gill P *et al*: **A high observed substitution rate in the human mitochondrial DNA control region**. *Nat Genet* 1997, **15**(4):363-368.

Table 1. Genetic equidistance explained by the maximum genetic diversity hypothesis. A hypothetical protein sequence of 10 amino acids is listed for each organism. Conserved positions are represented by numbers. Positions that change from time to time are represented by X. The hypothetical maximum diversity allowed for amphibian is 60%, for mouse 40%, and for human 10%.

Species	Sequence
Amphibian 1	0123xxxxxx
Amphibian 2	012326xxxx
Amphibian 3	012326xxxx
Mouse 1	012334xxxx
Mouse 2	01233424xx
Mouse 3	01233424xx
Human 1	012334315x
Human 2	012334315x
Human 3	012334315x

Maximum diversity (percent difference)	
Amphibian 1 *vs.* amphibian 2	60
Mouse 1 *vs.* mouse 2	40
Human 1 *vs.* human 2	10

Maximum distance (percent difference)	
Human *vs.* mouse	40
Human *vs.* amphibian	60
Mouse *vs.* amphibian	60

Table 2. Molecular clocks give consistent timing for macroevolution but inconsistent timing for microevolution. Percent identities between species are listed for four randomly selected genes. All four genes behave as good clocks in macroevolution from fish (*D. rerio)* to frog *(X. laevis)* to bird *(G. gallus)* to mouse (*M. musculus)* to human (*H. sapiens)*, which is consistent with the timing based on the fossil record as indicated for each divergence. In contrast, they give wildly contradictory timing when used to time microevolution divergence between pufferfish and zebrafish. The estimated time varies from 420 to 91 million years depending on which of the four genes is used as clock. The mutation rate or clock rate of each gene was derived from plotting the number of amino acid changes between protein sequences against species age estimated from fossil evidence. MyBP, million years before present. N.A., gene sequence not available.

	Percent identity				MyBP
	Prdm2	BTK	CytC	GCA1A	
H. sapiens v.s. *D. rerio*	39	61	80	66	450
H. sapiens v.s. *X. laevis*	55	N.A.	85	75	360
H. sapiens v.s. *G. gallus*	71	85	87	81	310
H. sapiens v.s. *M. musculus*	91	98	91	91	91
F. rubripes v.s. D. rerio	46				420
		71			400
			89		200
				91	91

Table 3. Genetic distance among different species of fungi. Three proteins, Pin1, CytC, and CMD from the baker's yeast were used to BLAST against the fungi database of NCBI . Percent identities in protein sequence between species of different genus, families, subphyla, and phyla are listed.

	Percent identity		
	Pin1	CytC	Cmd
Between genera within the same family *Saccharomycetaceae*			
S. cerevisiae v.s. D. hansenii/Debaryomyces	44	78	63
S. cerevisiae v.s. E. gossypii/Eremothecium	63		95
S. cerevisiae v.s. K. lactis/Kluyveromyces	68	84	94
Between families within the same order *Saccharomycetales*			
S. cerevisiae vs Y. lipolytica/Dipodascaceae	39	73	56
S. cerevisiae v.s. C. albicans/mitosporic Saccharomycetaceae	42	84	60
Between subphyla within the same phylum *Ascomycota*			
S. cerevisiae v.s. G. zeae PH-1/Pezizomycotina	42	67	
S. cerevisiae v.s. S. pombe/Schizosaccharomycetes	45	70	56
Between phyla within the same kingdom *Fungi*			
S. cerevisiae vs. R. oryzae/Zygomycota	43		
S. cerevisiae vs. C. Neo-formans/Basidiomycota	41	66	59
S. cerevisiae vs. C. cinerea/Basidiomycota	75	60	
S. cerevisiae vs. U. maydis/Basidiomycota	70	60	
S. cerevisiae vs. B. emersonii/Chytridiomycota			*58*

Table 4. The genetic distance between two fishes in many chromatin modifying enzymes is similar to that between a fish and a mammal. The percent identity between zebrafish (*D. rerio*) and pufferfish (*T. nigroviridis*), human (*H. sapiens*), or mouse (*M. musculus*) is shown for a number of chromatin modifying epigenetic enzymes. Genes are considered as having reached maximum distance in fishes if the distance between the two fishes is equal or slightly greater than between a fish and a mammal.

	(% identity) *D. rerio vs*.		
	T. nigroviridis	*H. sapiens*	*M. musculus*
Genes reached maximum distance			
Suv39H1/KMT1A	61	63	62
Smyd2/KMT3C	70	75	70
SET7/9/KMT7	71	73	73
PRDM11	61		64
PRDM4	57	59	59
PRDM15	60	63	63
PRMT4	81	81	85
Lsd1/KDM1	87	92	89
Jarid1b/KDM5b	62	62	62
MYST1/KAT8	87	87	85
SIRT5	71	75	71
HDAC1	80	83	82
HDAC4	78	77	79
Genes not yet reached maximum distance			
Suv4-20H1/KMT5B	59	53	54
EZH2/KMT6	82	77	76
PRDM2/KMT8	48	41	43
PRMT6	67	54	55
PRMT7	69	62	61
PRMT5	79	78	78
PRMT8	90	88	88
Jmjd2b/KDM4b	60	52	51
HAT1/KAT1	77	70	70
PCAF/KAT2B	88	82	78
CBP/KAT3A	66	61	61
MYST2/KAT7	89	77	77
Clock/KAT13D	73	70	69
SIRT3	66	55	58
SIRT4	73	64	66
SIRT6	76	73	72
SIRT7	63	55	54
HDAC3	96	92	92
HDAC8	84	73	75

Figure 1. Macroevolution and microevolution. The vertical direction is macroevolution and involves major changes in epigenetic complexity over time. The horizontal direction is microevolution and involves changes in varieties within a specific level of epigenetic complexity. The estimated number of species for each kind of organisms is indicated in parentheses. Time is not to scale and in the direction from past to future.

Figure 2. Genetic distance between two splitting organisms at various times during macroevolution. A 10 amino acid peptide with amino acids represented by numbers is shown to illustrate the dissimilarity or genetic distance between the species at various times during evolution. X represents amino acid positions that may change from time to time. A fraction of these X residues may be shared in different organisms due to common external environments that may differ from time to time. The ancestor organism A0 gives rise to two descendant lineages that gradually accumulate genetic distance until reaching a maximum at time T1. At some time point after the divergence, a punctuational jump in epigenetic complexity occurs in one of the lineages generating B1. The descendant organism A1 remains phenotypically similar to the ancestor A0. The lineage leading to B1 is phenotypically similar to A1 prior to the punctuational jump. The epigenetic jump in B1 reduces the genetic diversity of B1, as indicated by the reduction in the number of X positions.

Figure 3. The Neo-Darwinian hypothesis versus the maximum genetic diversity hypothesis. (A) The Neo-Darwinian model of microevolution and macroevolution. Genetic distance increases with time with no maximum cap. Fish and amphibians are used as examples. The transition from fish to amphibian is indicated by the dashed line. The starting point of the dashed line represents the time when amphibian epigenotype or phenotype first became obviously distinct from that of fish. **(B)** Model of microevolution and macroevolution based on the MGD hypothesis.

Figure 4. Inverse relationship between genetic diversity and epigenetic complexity. (**A)** Maximum genetic diversity within each type of organisms in cytochrome c correlates with the time since the first appearance of each type. The percent amino acid change in cytochrome c within each type of organism was obtained by BLAST against protein database at the National Center for Biotechnology Information. (**B)** High epigenetic complexity as measured by the number of cell types per organism inversely correlates with genetic diversity and the time since the first appearance of each organism. The first eukaryote is more complex than bacteria in having more cellular compartments and more epigenetic enzymes. The number of cell types is estimated based on the complexity of the nervous systems to be relatively the most in the first primate, less in the first mammals, and still less in the first vertebrate. The figure is meant to show this relative trend but does not intend to show the precise number of cell types.

Figure 5. Genetic distance between organisms at various times during macroevolution. **A.** Genetic distance according to the Neo-Darwinian gradual mutation hypothesis. **B.** Genetic distance according to the MGD hypothesis. A 10 amino acid peptide with amino acids represented by numbers is shown to illustrate the dissimilarity or genetic distance between the species at various times during evolution. X represents amino acid positions that may change from time to time. A fraction of these X residues may be shared in different organisms due to common external environments that may differ from time to time.

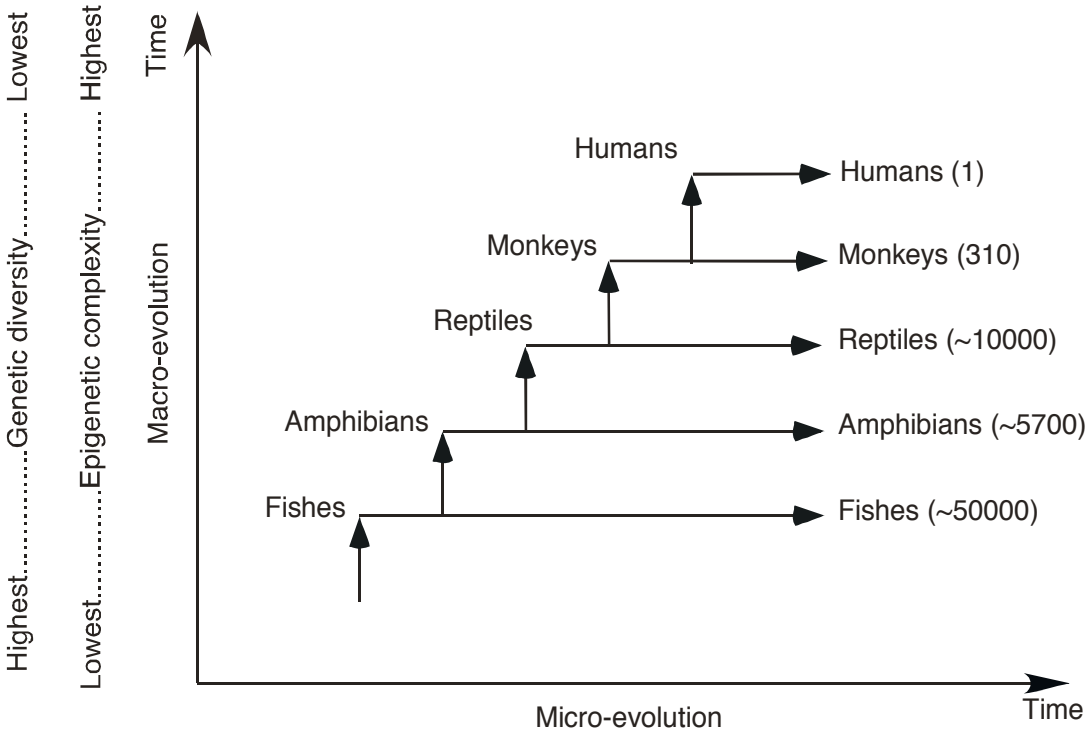
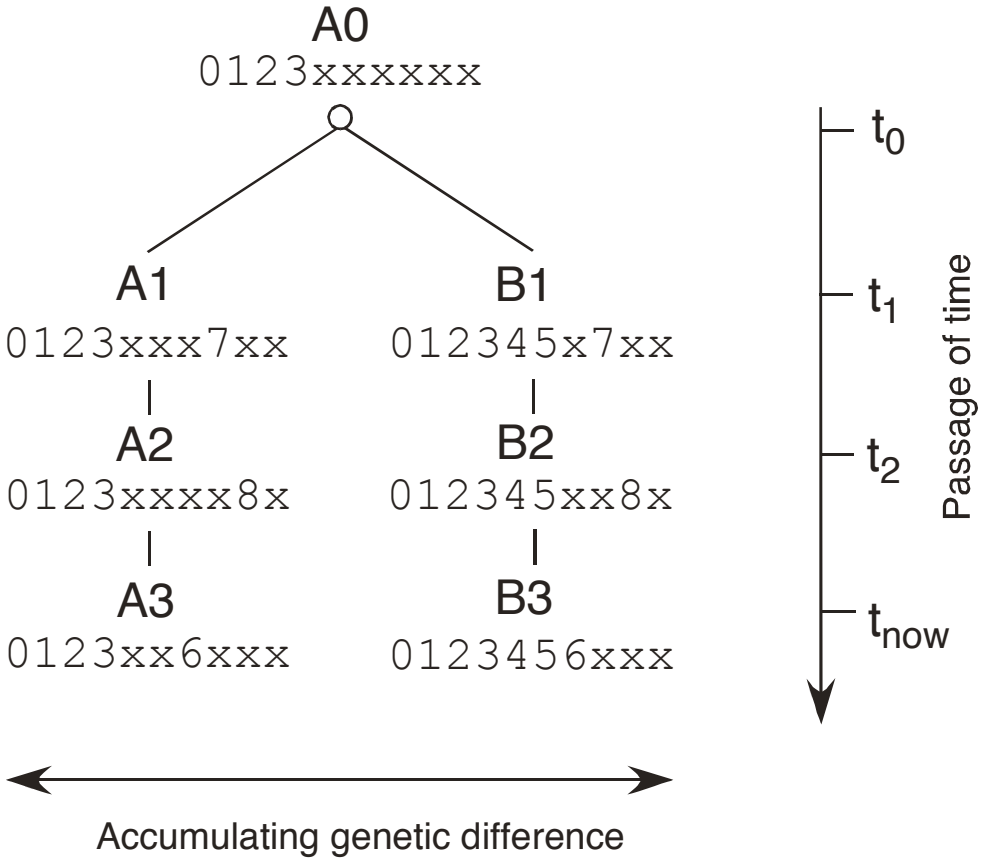


Figure 1

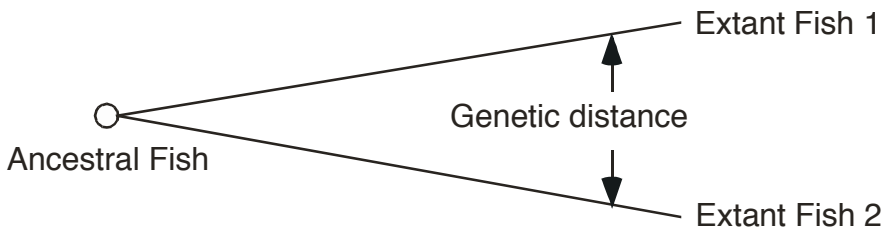


Distance (A1-B1) = Distance (A3-B3) = 50% dissimilarity
Distance (A1-B2) = 60% dissimilarity > Distance (A3-B3)
Distance (A2-B3) = 60% dissimilarity > Distance (A3-B3)
Distance (A3-B3) = Maximum distance within A3 (A3.1-A3.2)

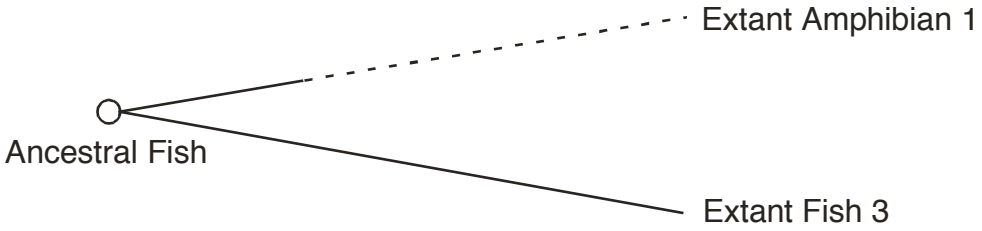
Figure 2

A

Neo-Darwinian microevolution model

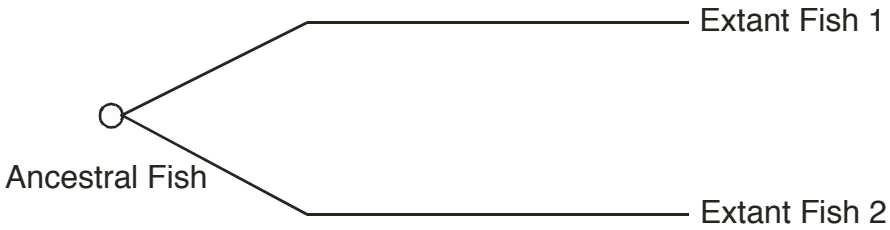


Neo-Darwinian macroevolution model



B

MGD microevolution model



MGD macroevolution model

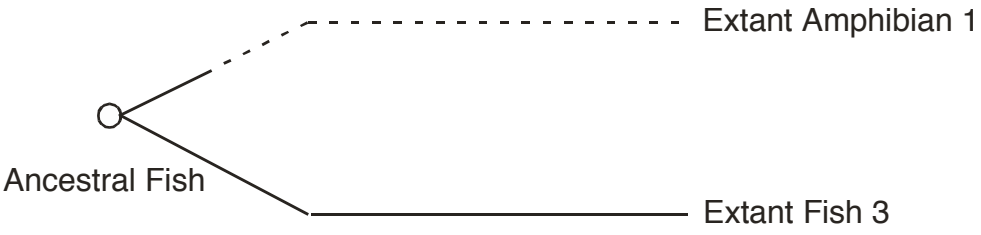
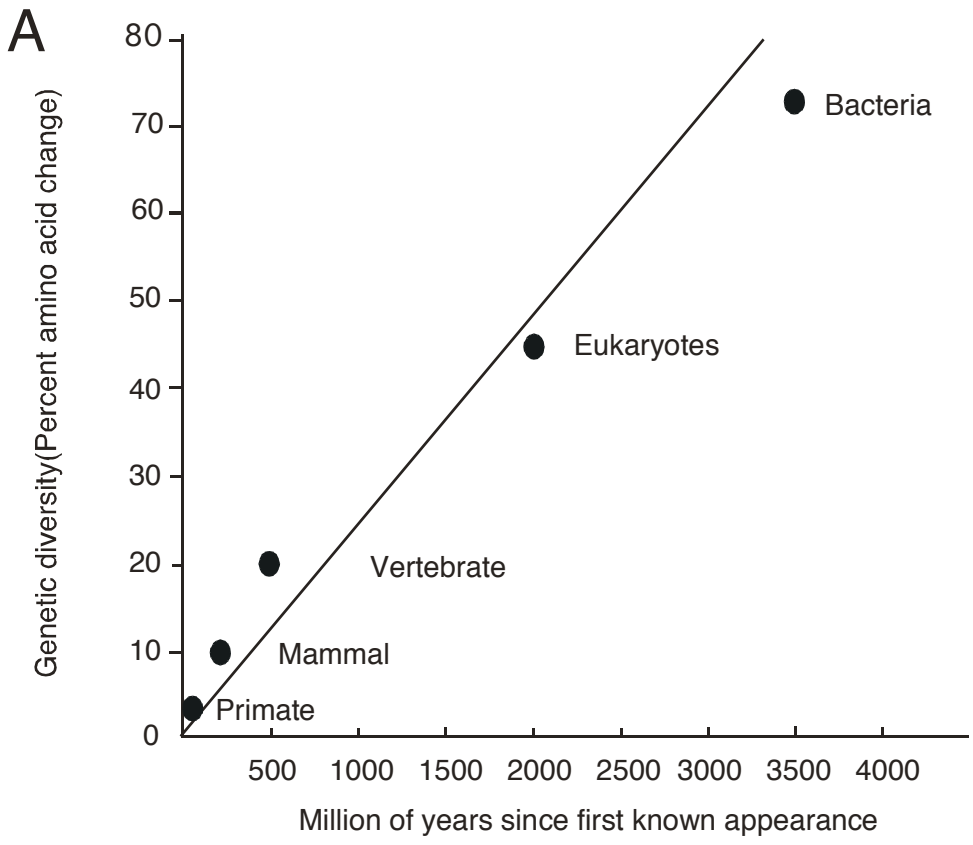
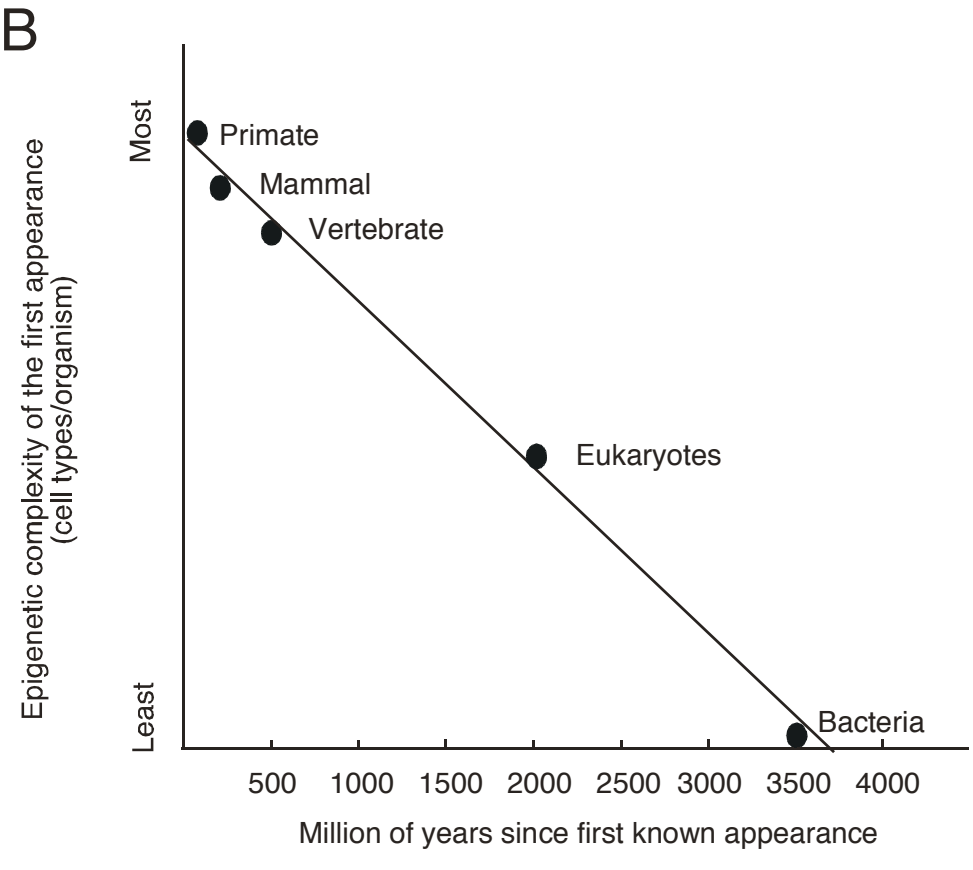


Figure 3

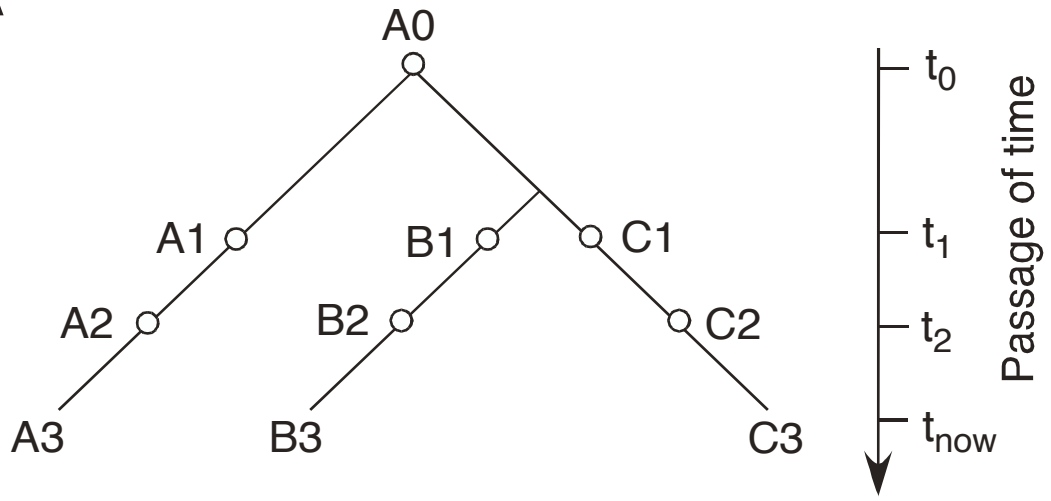




Least ..... Genetic diversity.....Most

Figure 4

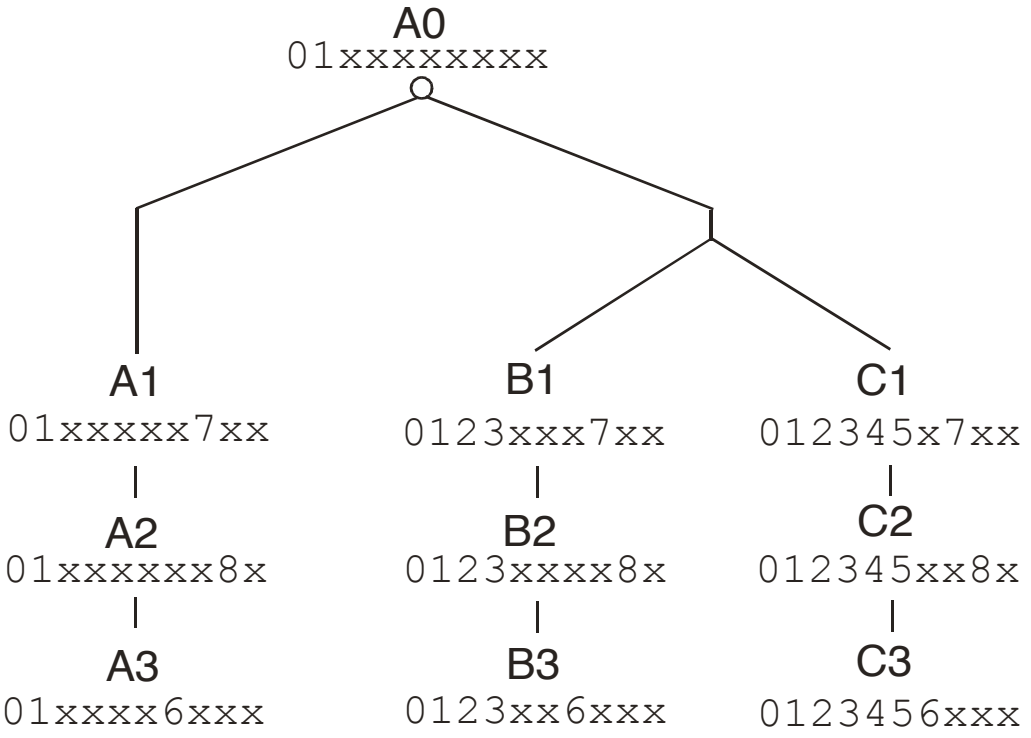
A



Distance (C1-B2) < Distance (C3-B3)

Distance (C1-A3) < Distance (B3-A3)

B



Distance (C1-B2) > Distance (C3-B3)

Distance (C1-A3) > Distance (B3-A3)

Figure 5

Supplementary Information to accompany

Inverse relationship between genetic diversity and epigenetic complexity

Shi Huang
The Burnham Institute for Medical Research
10901 North Torrey Pines Roads
La Jolla, CA 92037
shuangtheman at yahoo.com

Additional facts explained by the MGD hypothesis

1. *More recently evolved complex brachiopods are closer to mammals.* Here is another example of genetic non-equidistance to a more complex outgroup. The inarticulate brachiopod genus Lingula (order Linguilida) is the oldest, relatively evolutionarily unchanged animal known. The oldest Lingula fossils are found in Lower Cambrian rocks dating to roughly 550 MyBP. Terebratulids are modern articulate brachiopods and appeared later in evolution around ~430 MyBP. The molecular clock hypothesis predicts that mammals should be equidistant to Lingula and terebratulids. But the MGD hypothesis predicts that mammals should be closer to terebratulids given that they evolved later and should have lower genetic diversity. Indeed, a random sampling of several proteins showed that mammals are closer to terebratulids than to Lingula (Cox1, Cox2, Cox3, ND1, and COB). Also, terebratulids are closer to mammals than to a fellow brachiopod Lingula.

In contrast to the brachiopods, complex plants (flowering plants) that appeared later in evolution and simpler plants (mosses) that appeared earlier are about equidistant to mammals in several randomly analyzed genes (EF1a, Adh1a, EIF2b, Pin1, PP1, RPC1, and Cox1). The identity between flowering plants and mosses are much greater than between mammals and mosses, in contrast to brachopods where the distance between mammals and Lingula is similar to that between terebratulids and Lingula. Thus, plants have evolved plant-specific conserved domains since separating from mammals but before divergence of mosses and flowering plants.

2. *Radiation of mammals and the Cambrian explosion.* The two main areas of disagreement between molecular clocks and the animal fossil record concern the radiation of mammal orders around the Cretaceous-Tertiary boundary (65 MyBP) and animal phyla at the Cambrian explosion 520 MyBP [1]. In each case, molecular clocks show much deeper divergence. The MGD hypothesis suggests that the rates of change in genetic distance for macroevolution are determined by epigenetic complexity. They tend to be slower than the actual mutation rates. If these slower rates are used to date microevolution (most mammals may share similar level of epigenetic complexity), we would expect to see a deeper time of divergence than the actual time, as we have already seen above for the two fishes. For any estimation of divergence time, we must use slow evolving genes that have not reached maximum distance. But most studies today used fast evolving genes that have already reached maximum distance. Dating of the radiation of mammal orders with slowing evolving genes indeed showed results consistent with the fossil record (Huang, manuscript in preparation).

The rate of change in epigenetic programs between phyla may be much greater than that between different species within one phyla. For example, vertebrates have a much greater number of PRDM epigenetic enzymes than arthropods [2]. But the number of PRDM genes among different species of vertebrates is similar. The rate of change in epigenetic programs in macroevolution within the vertebrate phyla may be slower than that between phyla or between arthropods and vertebrates. So when the slow rate estimated from speciation events within one phyla, that of vertebrate, is used to calibrate the time of phyla divergence between arthropods and vertebrates, the time would be estimated to be deeper than the actual time (1000 MyBP versus 520 MyBP) [1].

3. *Actual mutation rate in real time is faster than that calculated from phylogenetic analysis*. It is well known that mutation rate from pedigree analysis on genealogical timescales is often an order of magnitude or more greater than mutation rate from phylogenetic analysis over geological time [3, 4]. Thus, phylogenetic diversity or distance over geological time is uncoupled from actual mutation rate observed on genealogical timescales. It suggests that actual mutation rates are often fast enough for most organisms to reach a maximum cap in genetic distance over geological timescale. Indeed, if actual mutation rates are slower than those from phylogenetic analysis, it would falsify the MGD hypothesis.

The phylogenetic diversity or distance reflects the maximum diversity allowed for an organism. Some of the variants at a particular time period accumulated as a result of random mutations may not persist long over geological time and may have to be replaced by another set of variants at a later time period (Figure 2). Maximum genetic distance between two species would stay constant over time while the same genetic distance may be maintained by different sets of variants at different times (Figure 2). A set of variants best suited for life at one time may not be the best at a different time and would have to be replaced.

4. *Stasis and punctuation in the fossil record.* The MGD hypothesis suggests that morphological phenotypes for complex organisms are better correlated with epigenotypes. Advances in epigenotypes in macroevolution occur largely via punctuation (Figure 2). Such punctuation events are followed by stasis in epigenotypes in microevolution. Thus the hypothesis predicts both stasis and punctuation at the level of epigenotypes and in turn at morphological levels. Consistently, the fossil record shows both stasis and punctuation at morphological levels [5].

5. *Copy number variations of the genome.* Advances in epigenetic complexity may involve changes that affect large regions of the genome, such as amplification or deletion of long stretches of DNA. Thus, such copy number changes may be expected to be a common behavior of the genome just like point mutations are. Indeed, copy number variations are observed to be common in the human genome [6]. Within a specific level of epigenetic complexity, a certain range of neutral and random copy number changes are allowed that may affect slightly epigenetic programs, just like a certain range of random point mutations are allowed. Relaxation of epigenetic programs is

expected to allow more abnormal copy number changes to occur. Indeed, cancer is commonly caused by loss of epigenetic control and often exhibits aneuploidy and amplifications or deletions of long stretch of DNA.

6. *Inverse correlation between genome size and genetic diversity*. Large size genomes (measured here as number of genes) require more complex epigenetic regulation than small genomes and are expected to show less genetic diversity. Indeed, there is a strong inverse correlation between genome size and genetic diversity [7]. Genetic diversity is more responsive to changes in genome size in bacteria than in eukaryotes, indicating that genetic diversity is restricted more by epigenetic complexity than by genome size in eukaryotes.

In microbes, there is an inverse relationship between genome size and mutation rate per base pair per replication [8]. In four metazoans analyzed, the mutation rate per base pair per replication is lowest for humans, higher for mice, and still higher for drosophila or worm. These data are expected from the MGD hypothesis.

7. *No bacterium lineage could be identified as the closest relative of eukaryotes.* Based on the overall trend in evolution from simple to complex organisms and the earliest fossil evidence of life on Earth, it is almost certain that bacteria were the ancestors of the eukaryotes. However, the MGD hypothesis predicts that no single bacterium lineage could be identified among bacteria as the closest relative of eukaryotes. Such a lineage, if indeed exists, would have long reached maximum diversity and would show equidistance to eukaryotes as other bacteria. In contrast, if there is no maximum cap on diversity or if time is not long enough yet and if the Neo-Darwinian hypothesis is true, one should be able to identify the bacterium lineage that is closer to eukaryotes than most other bacteria. But extensive studies show that no such bacterium lineage can be identified. Recent data show that the identification of archaea as closer to eukaryotes is only true for some class of genes such as those involved in translation [9, 10]. For many other genes, archaea are in fact more distant to eukaryotes than eubacteria. The overall pattern of genetic similarity suggests that common selection and coincidence may account for most of the sequence identities between eukaryotes and bacteria.

The closer relationship between a bacterium species and eukaryotes in some genes but not others has been commonly interpreted to mean horizontal gene transfer, even though there is little independent evidence for it. It is more likely however that the closer relationship are fortuitous due simply to the fact that bacteria have much greater genetic diversity and some gene variants of bacteria would by chance resemble an eukaryotic version. If one compares a gene from a mammalian species against orthologous genes of all species of bacteria in the Genbank, one would find that the degree of similarity would vary to a great extent (e.g., for GLUD1, the identity between human and all bacteria ranges from 30% to 50%). In contrast, if one compares a gene from an individual bacterium species against all vertebrate species in the Genbank, one would find that the degree of similarity falls within a very narrow range (e.g., for GLUD1, the identity between the bacterium *Pedobacter sp. BAL39* and all vertebrates ranges from 47% to 53%). Vertebrates have lower genetic diversity and there is less probability for a variant of vertebrates to be more closely related by chance than other variants to an individual variant of bacteria.

There are data against the idea of horizontal gene transfer. If a gene was transferred from a prokaryotic lineage into the vertebrate lineage, this likely occurred within the past 400 to 500 million years, after most of the major prokaryotic phyla were established. Therefore, any transferred gene should be more closely related to its donor lineage than to any other prokaryotic lineage, which would be detectable in phylogenetic trees. However, it was found that most of the genes shared between vertebrates and bacteria did not show patterns consistent with bacterial to vertebrate gene transfer [11].

8. *Stability of epigenetic programs.* It is well known that artificial selection or breeding of animals can only generate varieties of the same type but never of a different type. This fact plainly indicates that genetic variation within an organism is not without a limit. The epigenetic program that allows a genome to manifest a dog phenotype also prevents the same genome from randomly drifting into something that is not allowed by the epigenetic program. Indeed, random drifting is far more likely to give rise to cancer rather than a novel functional organ. If genotypes can be rather unstable or easily influenced by mutations, the epigenotypes are relatively much more stable. Indeed, when cultured for up to *ten years*, hundreds of cell divisions later, *Drosophila* wing disc cells can still give rise to adult wing structures [12]. The stability of epigenotypes is also indicated by the stasis and extinction phenomena in the fossil record. If environment becomes unsuitable for survival, a species would more often than not go extinction rather than change itself in its epigenetic programs. In today's world, all we observe is extinction of species rather than drastic transformation of species.

A specific epigenetic program allows a certain degree of variation in genotypes and in turn a certain range of adaptive capability in response to environmental changes. When the environmental changes exceed the adaptive capability allowed within a specific epigenetic program, the organism would simply go extinct rather than change. Change is not without a limit. Change is not the only feature of evolution. Equally important and obvious as change is the opposite of change. If constant random change within a limited range in genotypes is a hallmark of evolution, then long period of stasis and stability in epigenotypes followed by short period of punctuational advance in epigenotypes is an equally important hallmark of evolution. Indeed, the genetic code is the optimal code for error minimization or for minimizing the effects of random changes; it is the most stable of all possible codes and is optimal for stability rather than for random changes [13].

9. *Low genetic diversity of chromosome X.* Comparisons of genomes have shown a lower rate of sequence divergence on chromosome X than the autosomes for many species. Humans and chimpanzees are far closer in X than in autosomes [14]. Chromosome X undergoes X inactivation in females, which is an epigenetic event. Thus, genes located on X encounter more epigenetic restrictions than genes on autosomes and are therefore expected from the hypothesis to be less tolerant of mutations. This explanation is far more reasonable than the

suggestion of interbreeding between humans and chimpanzees [14].

10. *The genetic diversity of tuatara*. The tuatara of New Zealand is a living fossil reptile and has very slow metabolic and growth rates, long generation times and slow rates of reproduction. Contrary to expectations from Neo-Darwinian theory, tuatara has high 'mutation rates', significantly higher than those of mammals measured in real time by the same method of using mitochondrial D-loop DNA sequences from fossils of ~10,000 years old [15]. However, this result is to be expected from the MGD hypothesis since reptiles should have higher maximum genetic diversity than mammals. If ~10,000 years is sufficient for reptiles and mammals to reach maximum cap in genetic diversity in the D-loop region, then the reptiles would show higher genetic diversity, resulting in the appearance of a higher 'mutation rate'. But in reality, tuatara can have slower mutation rate but still show higher genetic diversity than a mammal if time is long enough for tuatara to reach the maximum cap or to be close to the cap.

References

1. Hedges SB: **The origin and evolution of model organisms**. *Nat Rev Genet* 2002, **3**(11):838-849.
2. Huang S: **Histone methyltransferases, diet nutrients, and tumor suppressors**. *Nat Rev Cancer* 2002, **2**:469-476.
3. Ho SYW, Larson G: **Molecular clocks: when times are a-changin'**. *Trends Genet* 2006, **22**:79-83.
4. Millar CD, Dodd A, Anderson J, Gibb GC, Ritchie PA, Baroni C, Woodhams MD, Hendy MD, Lambert DM: **Mutation and evolutionary rates in adelie penguins from the antarctic**. *PLoS Genet* 2008, **4**(10):e1000209.
5. Gould SJ, Eldredge N: **Punctuated equilibrium comes of age**. *Nature* 1993, **366**(6452):223-227.
6. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W *et al*: **Global variation in copy number in the human genome**. *Nature* 2006, **444**(7118):444-454.
7. Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P: **Toward automatic reconstruction of a highly resolved tree of life**. *Science* 2006, **311**(5765):1283-1287.
8. Drake JW, Charlesworth B, Charlesworth D, Crow JF: **Rates of spontaneous mutation**. *Genetics* 1998, **148**(4):1667-1686.
9. Pennisi E: **Genome data shake tree of life**. *Science* 1998, **280**(5364):672-674.
10. Britten RJ: **Idiosyncratic evolution of conserved eukaryote proteins that are similar in sequence to archaeal or bacterial proteins**. 2008:Available from Nature Precedings.
11. Salzberg SL, White O, Peterson J, Eisen JA: **Microbial genes in the human genome: lateral transfer or gene loss?** *Science* 2001, **292**(5523):1903-1906.
12. Hadorn E: **Dynamics of determination**. *Symp Dev Biol* 1967, **25**:83.
13. Freeland SJ, Knight RD, Landweber LF, Hurst LD: **Early fixation of an optimal genetic code**. *Mol Biol Evol* 2000, **17**(4):511-518.
14. Patterson N, Richter DJ, Gnerre S, Lander ES, Reich D: **Genetic evidence for complex speciation of humans and chimpanzees**. *Nature* 2006, **441**(7097):1103-1108.
15. Hay JM, Subramanian S, Millar CD, Mohandesan E, Lambert DM: **Rapid molecular evolution in a living fossil**. *Trends Genet* 2008, **24**(3):106-109.