The Most Primitive Extant Ancestor of Organisms and Discovery of Definitive Evolutionary Equations Based on Complete Genome Structures

Kenji Sorimachi[1,2]

[1]Educational Support Center, Dokkyo Medical University, Mibu, Tochigi, Japan; [2]Bioresearch Laboratory, Gunma Agriculture and Forest Development, Co. Ltd., Takasaki, Gunma, Japan

Correspondence to: Kenji Sorimachi, kenjis@dokkyomed.ac.jp

ABSTRACT

Evolutionary divergence has been characterized based on morphological and molecular features using rationale based on Darwin's theory of natural selection. However, universal rules that govern genome evolution have not been identified. Here, a simple, innovative approach has been developed to evaluate biological evolution initiating the origin of life: whole genomes were divided into several fragments, and then differences in normalized nucleotide content between nucleotide pairs were compared. Intramolecular nucleotide differences in complete mitochondrial genomes reflect evolutionary divergence. The values of $(G - C)$, $(G - T)$, $(G - A)$, $(C - T)$, $(C - A)$ and $(T - A)$ reflect biological evolution, and these values except for $(G - C)$ and $(T - A)$ change inversely to positive from negative along biological evolution of bacterial genomes. More highly evolved organisms, such as primates and birds, seem to have greater levels of $(C - T)$ in mitochondria. Based on nucleotide content structures, *Monosiga brevicollis* mitochondria may be the most primitive extant ancestor of the species examined here. The two normalized nucleotide contents are universally expressed by a linear regression line, $(X - Y)/(X + Y) = a(X - Y) + b$, where X and Y are nucleotide contents and (a) and (b) are constants. The value of $(G + C)$, $(G + A)$, $(G + T)$, $(C + A)$, $(C + T)$ and $(A + T)$ was ~0.5. Plotting $(X - Y)/(X + Y)$ against X/Y showed a logarithmic function $(X - Y)/(X + Y) = a \ln X/Y + b$, where (a) and (b) are constant. Nucleotide content changes are expressed by a definitive equation, $(X - Y) \approx 0.25 \ln(X/Y)$.

1. INTRODUCTION

Primitive organisms might appear after long periods of chemical evolution, during which various or-

---

ganic compounds were accumulated. Recently, it was reported that vesicles consisting of lipids and poly-nucleotides spontaneously replicated under experimental conditions [1]. These pseudo-organisms were the first living things to appear on Earth; however, we cannot trace their origins because the original assembly of chemical compounds was not stable. Only those primitive organisms that established a method of self-replication could survive and continue to evolve. Therefore, it is difficult to identify the true primitive ancestors of life on Earth, as their very make-up would probably be unstable in the current environmental conditions. Although fossilized traces of very early organisms have been found in sedimentary rocks dating from 3.1 - 3.7 billion years ago [2-5], they do not appear to be the origin of life. It is likely that only primitive organisms that had cell walls could leave a fossil record, which may rule out finding any traces of simpler organisms.

Several key eukaryotic organelles originated from symbioses between separate single-celled organisms [6, 7]. For example, mitochondria developed from the proteobacterium *Rickettsia* or its relatives [8, 9]. The *Reclinomonas americana* (Protist) mitochondrion (~70 kb), consisting of 97 genes, is thought to be an ancestral mitochondrial DNA (mtDNA), while vertebrate mtDNA (~16 kb), consisting of 37 genes, seems to be constructed with only essential genes for respiration reactions. Based on 13 respiratory genes, the amino acid composition and gene patterns within complete mitochondrial genomes are almost identical among all animal species, except Amoebazoa [10]. Irreversible evolutionary divergence accompanied by increasing G and C content means that the G and C content of descendants should be identical to or higher than that of the ancestor. However, it was shown that the G and C values for *R. americana* (G: 0.148, C: 0.114) mtDNA were higher than those of *Monosiga brevicollis* (G: 0.081, C: 0.059), which were the lowest among the samples examined [11]. In addition, nucleotide regression lines for the two species differed from each other [11, 12]. Thus, we concluded that mitochondria might derive directly from primitive organisms [13]. Our present results based on genome evolutionary rules deny that the *Reclinomonas americana* mitochondrion is thought to be an ancestral mitochondrial DNA [8, 9]. However, there is no scientific evidence that mitochondria are the most primitive extant ancestor of all life. Based on Charles Darwin's theory of natural selection, all organisms have a single origin and common ancestor, therefore, we predict that the amino acid compositions of their complete genomes should naturally reflect their evolution from bacteria to *Homo sapiens* [14, 15].

The two nucleotide relationships were expressed by linear regression lines, which crossed at a single point [11, 12]. Thus, we concluded that the origin of life was single and pluripotent. Assuming this conclusion, at the crossing point representing the origin of life, ($G \approx C$) and ($T \approx A$) to satisfy Chargaff's parity rules [16, 17]: ($G - C$) ≈ 0 and ($T - A$) ≈ 0. The two regression lines representing high and low C/G animal mitochondria must satisfy ($G \approx C$) and ($T \approx A$) at the crossing point for the line based on Chargaff's parity rule. Therefore, differences in ($G - C$) and ($T - A$) reflect mitochondrial evolution. In addition, to satisfy Chargaff's second parity rule [17], the nucleotide content of the first half of the DNA strand should be equal to that of the second half [18], providing a symmetry between the 5' and 3' ends. As animal mitochondrial genomes deviate from Chargaff's second parity rule [17], the divergence in animal mitochondria increases the unsymmetrical nucleotide content. To detect intragenomic alterations based on increases in the unsymmetrical nucleotide content, a method that investigates nucleotide content differences, *i.e.* ($G - C$), ($G - A$), ($G - T$), ($C - A$), ($C - T$) and ($A - T$) in the sequentially divided genomes, was developed in the present study. These exciting results could not have been obtained from sequence analyses.

2. MATERIALS AND METHODS

All genome sequences were obtained from the National Center for Biotechnology Information GenBank database (http://www.genome.jp/ja). The nucleotide contents of all genomes, which are listed in Extended Data Figs, were normalized ($G + C + A + T = 1$) because normalized values are independent of species and genome sizes [19, 20]. A whole genome was divided into several fragments to evaluate intra-genome alterations due to evolution, and the differences between the two nucleotide contents of six

pairs (($G - C$), ($G - T$), ($G - A$), ($C - T$), ($C - A$), ($T - A$)) were calculated for all fragments. In addition, mitochondria were divided into three equal-size fragments. To easily understand evolution at a glance, these six nucleotide differences were graphically expressed with a single pattern. This appears to be a novel approach. Calculations were carried out with a personal computer, TOSHIBA dynabook T552 (Tokyo, Japan), Windows 10 installed.

3. RESULTS AND DISCUSSION

3.1. Primitive Ancestor

Chargaff's parity rules, where ($G = C$) and ($T = A$), were the original rules for inter- [16] and intra- [17] molecular relationships of DNA, and are applicable to chromosomal DNA [21-23]. The normalized nucleotide contents (C, T and A), predicted from complete genomes, are expressed by the content of the fourth nucleotide (G) using linear regression lines in chromosomal [21], non-animal mitochondrial and chloroplast DNA [22, 23]. Classifying animal mitochondria into high and low C/G groups, the content of the first three nucleotides could be similarly expressed by the forth nucleotide content [23], although animal mitochondria deviated from Chargaff's second parity rule [24, 25] (**Figure 1(a)**). Vertebrate mitochondria overlapped with the high C/G invertebrates, which have high C content (**Figure 1(a)**), indicating that both groups were descended from the same origin [12, 13], and that more highly evolved organisms seem to have a greater cytosine content [11]. Non-animal mitochondria and chloroplasts obeyed Chargaff's rules [23]. C contents in cellular organelles and chromosomal DNA vs G contents were expressed with linear regression lines, which crossed at a single point (**Figure 1(a)**). This is consistent with our previous findings [11, 12]. In addition, we found that nucleotide content relationships in viral DNA (**Extended Data Table 1**) were heteroskedastic (**Figure 1(b)**).

Based on the normalization of the four nucleotides ($G + C + T + A = 1$), the GC content, ($G + C$), is expressed by $\{1 - (A + T)\}$, where GC content ($G + C$) and AT content ($A + T$) are completely linear, not only in chromosomal, non-animal mitochondrial, and chloroplast DNA, but also in animal mitochondrial DNA (**Figure 1(c)**). The same result was obtained for viral DNA (**Figure 1(d)**). Assuming irreversible divergence, it is generally thought that GC content increases along with biological evolution. Thus, the organelles that have the lowest GC content might be the most primitive. In the current study, the GC content of the mtDNA of the choanozoan *M. brevicollis* was the lowest among the samples examined. The GC content of the bacterium *Streptomyces coelicolor* was the highest of all the samples, although this organism was not the most evolved. Amongst the viruses examined, the GC content of *Melanoplus sanguinipes* Entomopoxvirus was the lowest, while that of *Papio ursinus* Cytomegalovirus was the highest (**Figure 1(d)**).

A previous study [11] stated that the C content of complete mitochondrial genomes reflects biological evolution better than the GC content. Based on the normalization of the four nucleotides, $G + C + A + T = 1$, $C = 1 - (G + A + T)$. Thus, C content and ($G + A + T$) are linear. In the current study, the lowest C content was observed in *M. brevicollis* mtDNA, while the highest C content was found in mtDNA from avian species *Gallus gallus* and *Taeniopygia guttata*, and primates *H. sapiens*, *Pan paniscus*, *Pan troglodytes*, and *Gorilla gorilla* (**Figure 1(e)**). Amongst the viral genomes, the C content of *Mollivirus sibericum* was the lowest, while that of De-Brazza's monkey virus was the highest (**Figure 1(f)**).

3.2. Genome Evolution

The complete genome is represented by four nucleotide contents based on more than a certain amount of randomly chosen fragments, as well as on completely linear fragments [26, 27]. Therefore, when the whole genome of bacterium *Ureaplasma urealyticum* (G; 0.131, C; 0.127) was sequentially divided into nine equal fragments, the amounts of the four nucleotides in each fragment were quite similar (**Figure 2(a)**). This is consistent with our previous results, which indicated that a whole genome may be constructed from small units with similar amino acid compositions [26, 27]. Nucleotide content differences (*i.e.* ($G - C$) and ($A - T$) were reversed from positive to negative values in fragments 5 and 6 of the *U. urealyticum* genome (**Figure 2(b)** and **Figure 2(c)**). The ratios of ($G - C$)/($G + C$) and
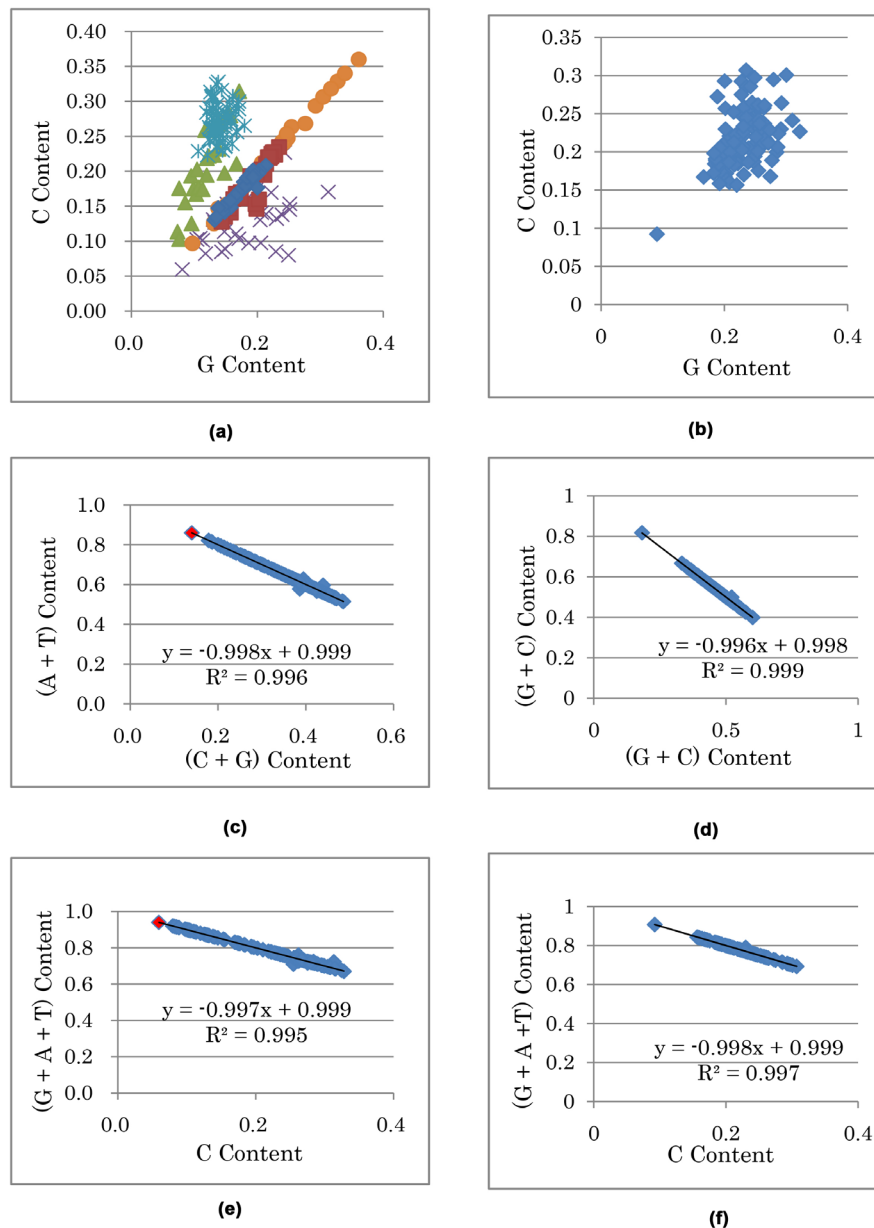
Figure 1. Regression lines. Left and right panels represent cellular organelles and viruses, respectively. Upper panels (a and b): horizontal and vertical axes represent G and C contents, respectively. Upper left panel: C content plotted against G content in vertebrate mitochondria (asterisk), high C/G invertebrate mitochondria (triangle), low C/G invertebrate mitochondria (cross), bacteria (circle), non-animal mitochondria, chloroplasts (diamond), and chromosomes (square). Red square represents *Monosiga brevicollis* mitochondrion. As the *p*-value based on the two regression lies representing high and low C/G animal mitochondria is 0.000, these two lines are significantly different each other. There is no-significant difference among chromosomal, non-animal mitochondrial and chloroplast DNA, while these lines differ from those representing animal mitochondrial DNA. Right upper panel (a): C content plotted against G content in viruses. Middle panels (c and d): Horizontal and vertical axes represent (G + C) and (A + T) contents, respectively. Lower panels (e and f): Horizontal and vertical axes represent (C) and (G + A + T) contents, respectively.
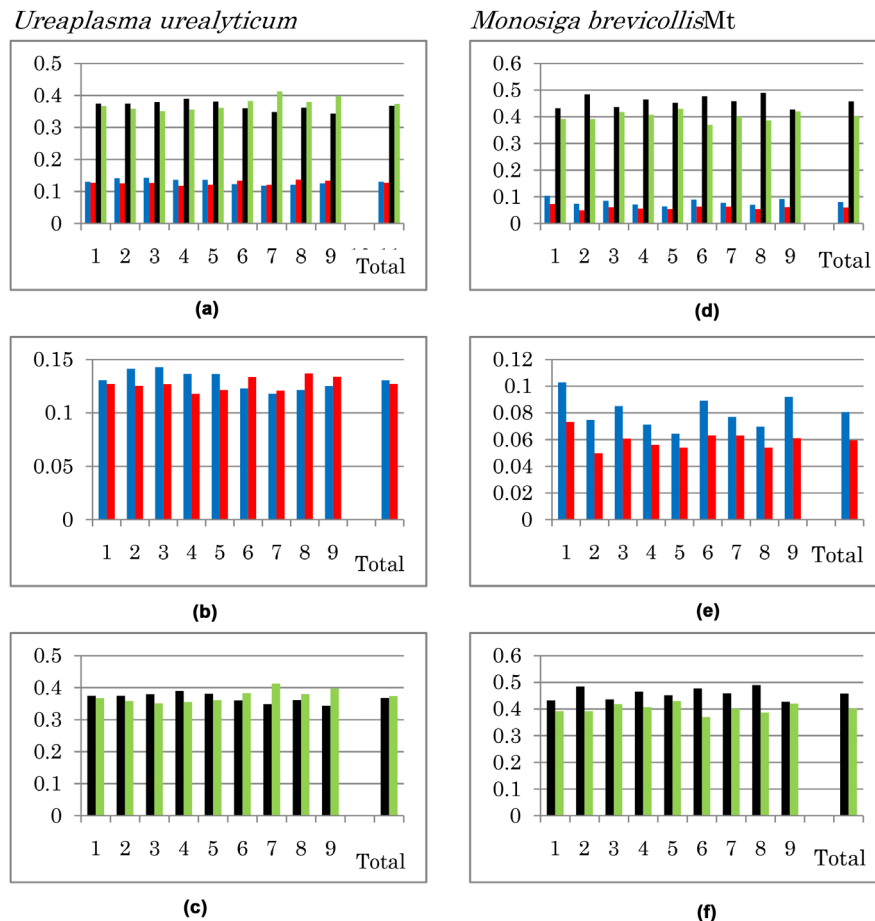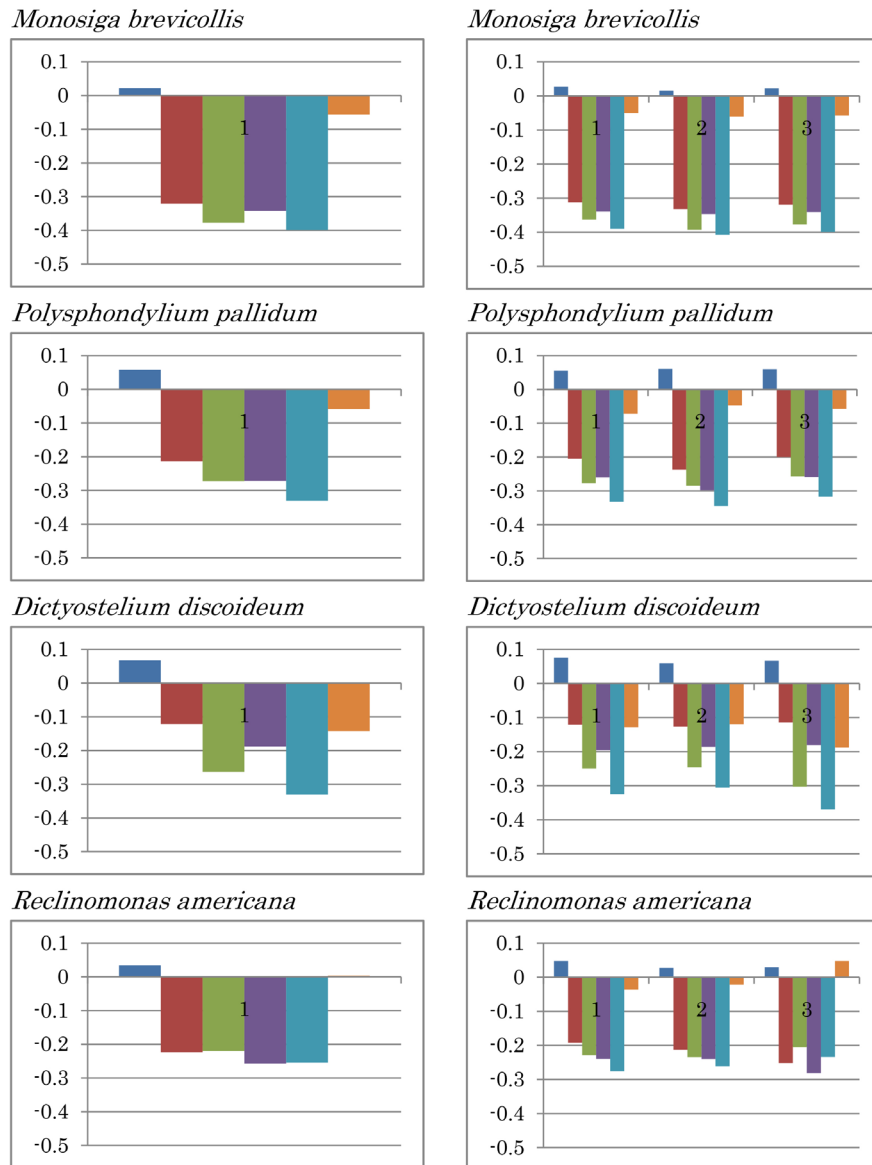
Figure 2. Nucleotide content was sequentially divided into nine fragments. Left side: *Ureaplasma urealyticum* mitochondrion, right side: *Monosiga brevicollis* mitochondrion. G: blue, C: red, A: black, and T: green. Vertical axis represents nucleotide content.

(A − T)/(A + T) are called the GC and AT skew, respectively [28]. The skew seems to be based on differences in replication processes between the leading and lagging strands [29]. In particular, replication of the lagging strand increases the probability of mutations as a result of the deamination of cytosine, and the inversion of nucleotide content differences reflects biological divergence. Similar phenomena are observed in mitochondria, which consist of heavy (H) and light (L) chains [30-32]. Plotting the GC skew vs. G content was used to classify animal mitochondria into two groups: high and low C/G [11]. In *M. brevicollis* mitochondria, the nine DNA fragments showed almost the same nucleotide contents (**Figure 2(d)**), as was observed in *U. urealyticum* (**Figure 2(a)**). However, GC and AT content difference inversions were not observed in *M. brevicollis* mtDNA (G: 0.081, C: 0.059) (**Figure 2(e)** and **Figure 2(f)**). Thus, the *M. brevicollis* mitochondrion might be more primitive than the *U. urealyticum* chromosome. These results clearly indicate that nucleotide content differences such as (G − C) and (A − T) reflect biological evolution. Therefore, the other nucleotide content differences, (G − A, G − T, C − A and C − T), were examined to determine whether or not these values reflect biological evolution.

3.3. Organelle Evolution

Complete mitochondrial genomes were investigated (**Figure 3**, left panel). To allow simple visual comparison of inter- and intra-species genome structures, genomes were sequential divided into three

Figure 3. Nucleotide content differences in complete mitochondrial genomes (left side) and the three sequential fragments of each mitochondrial genome from 1 to 9 (right side). Left to right: (G − C), (G − T), (G − A), (C − T), (C −A) and (T − A).

fragments throughout subsequent analyses, from which three separate patterns emerged. Six nucleotide content differences were observed among the mitochondria of the four species (*M. brevicollis*, *P. pallidum*, *D. discoidium* and *R. americana*) (Figure 3, right panels). The six nucleotide content patterns were conserved within the three fragments among the four species. No inversion of nucleotide content differences was observed in the mtDNA of *M. brevicollis* (G: 0.081, C: 0.059), the mycetozoan *Polysphondylium pallidum* (G: 0.143, C: 0.085), or *Dictyostelium discoideum* (G: 0.171, C: 0.104) (Figure 3), although differences in (G − C) and (T − A) values for *M. brevicollis* mtDNA were the lowest amongst these species. Based on genome sequencing, choanoflagellates are most closely related to animals [33]. As the nucleotide content difference patterns of the three fragments were almost identical for these three species, their nucleotide distributions were judged to be homogeneous, indicating nucleotide content symmetry. Thus, these

mitochondria are likely to be primitive. Consistent results were obtained from Ward's clustering analysis using amino acid compositions predicted from complete mitochondrial genomes as traits [11]. These findings indicate that the *M. brevicollis* mitochondrion is the most primitive among the three. In contrast, AT inversion was observed in the third fragment of *Reclinomonas americana* mtDNA (G: 0.148, C: 0.114), which has previously been proposed as a mitochondrial ancestor [8]. However, differences in (G − C) and (T − A) values in *R. americana* mtDNA were smaller than those in the mtDNA of the previous three organisms. Nucleotide content inversion causes significant differences in nucleotide content patterns as a result of unsymmetrical nucleotide content. Thus, the *R. americana* mitochondrion is probably more evolved than the former three mitochondria. In addition, AT inversion occurred in the following more highly evolved organisms: Mollusca species, squid (*Todarodes pacificus*), octopus (*Octopus vulgaris*), Echinodermata species, sea urchin (*Paracentrotus lividus*), water flea (*Daphnia pulex*), hermit crab (*Pagurus longicarpus*), and Humboldt squid (*Dosidicus gigas*) (Extended Data Figure 1). In addition, large positive (G − A) values in the three fragments were observed in *Paragonimus westermani*, while large positive (G − C) and (A − T) values in the three fragments were observed for the mtDNA of representatives of the following phyla: Cnidaria (*Pavona clavus*), Platyhelminthes (*Schistosoma mansoni*), Porifera (*Geodia neptuni*), Arthropoda (*Tigriopus californicus*), and Chordata (*Branchiostoma belcheri*) (Extended Data Figure 1).

A positive (C − T) value was characteristically observed in the three fragments of Echinodermata species *Acanthaster planci*, the second and third fragments of Mollusca species *Haliotis rubra*, and in the first fragment of Mollusca species *Lampsilis ornata*. CT inversion occurred in *H. rubra* and *L. ornata* mtDNA (Extended Data Figure 1). The nucleotide content difference patterns of the mtDNA of hemichordates *Saccoglossus kowalevskii* and *Balanoglossus carnosus* differed from each other. Both AT and CT inversions occurred in the first mtDNA fragment of *S. kowalevskii*, while large positive (C − T) and (C − A) value differences occurred in the second and third fragments. AC and AT inversions were observed in *B. carnosus* mtDNA (Extended Data Figure 1). Neither nucleotide inversion nor positive nucleotide differences were observed in the mtDNA of deuterostom, *Xenoturbella bocki*.

In the mtDNA of primate species *H. sapiens*, *P. troglodytes*, *G. gorilla*, *Macaca mulatta*, *Daubentonia madagascariensis*, *Nycticebus coucang*, and *Tupaia belangeri*, nucleotide content difference patterns were quite similar in the first four species, and large positive increases in (C − T) differences in the three fragments clearly indicated evolutionary divergence (Figure 4). The positive (C − T) differences in all three fragments were characteristic of these four primate mitochondria, while positive increases in (C − T) values were only observed in the third fragment of *N. coucang* and *T. belangeri* mtDNA. In contrast, nucleotide content difference patterns of the prosimian *Lemur catta* completely differed from those of the primates, although TA inversion was observed in the second fragment. The primate mtDNA nucleotide content patterns were also completely different from that of hemichordate *B. carnosus*, although their C contents were the highest among all organisms examined [11]. This finding indicates that mitochondrial structures respect epigenomic evolutionary functions.

The mitochondria of other vertebrates (Extended Data Text.), rodents (Extended Data Figure 2(a)), ocean-dwelling mammals, cetaceans, aves (Extended Data Figure 2(b)), amphibians (Extended Data Figure 2(c)), reptiles (Extended Data Figure 2(d)), and fishes (Extended Data Figure 2(e) and Extended Data Figure 2(f)) were also examined. In these organelles, differences in (G − C) and (A − T) or in other nucleotide inversions, as well as GC and AT inversions, increased along with evolution. Consistent results were also obtained from non-animal mitochondria and chloroplasts (Extended Data Figure 3), as well as prokaryotes (Extended Data Figure 4). In the microsporidian protozoan *Encephalitozoon cuniculi*, GC and AT inversions, as well as other nucleotide content inversions, were observed (Extended Data Figure 5).

3.4. Virus Evolution

The *M. sanguinipes* Entomopoxvirus genome had the lowest G content among the viruses examined (Figure 1), and AT inversion was also observed (Extended Data Figure 6). In *P. ursinus* Cytomegalovirus,
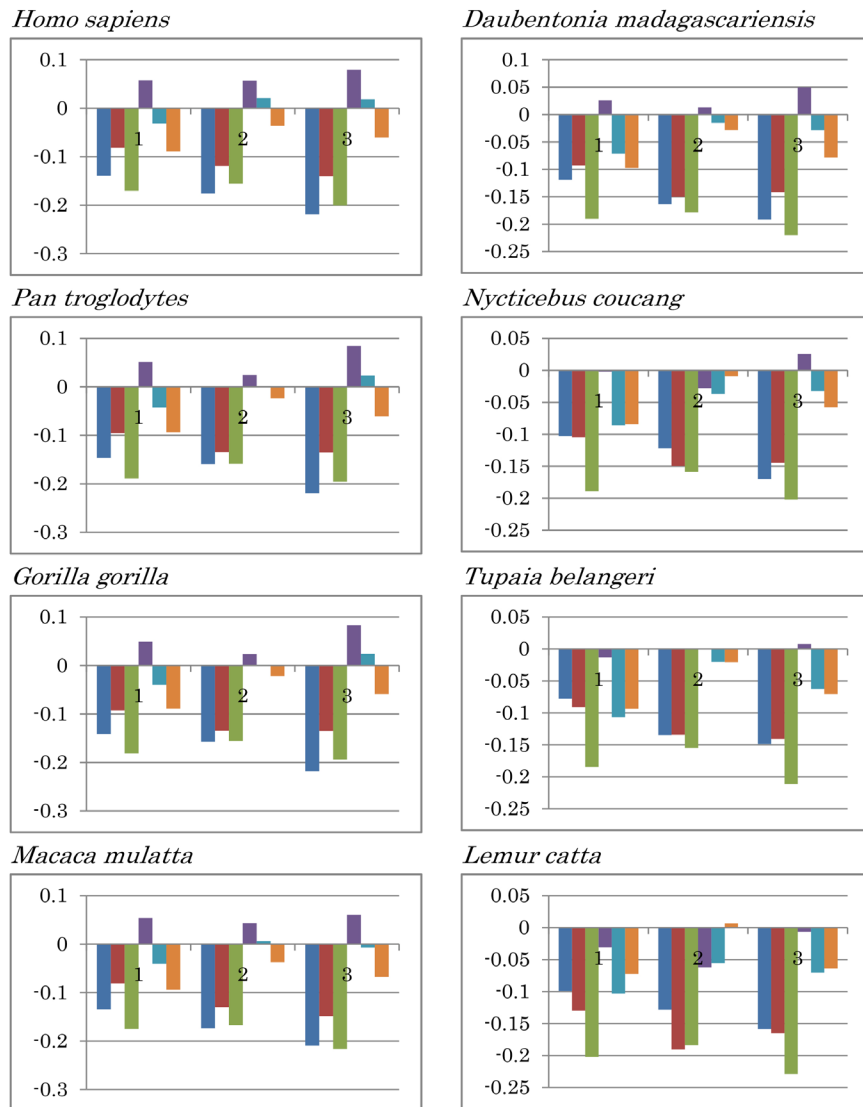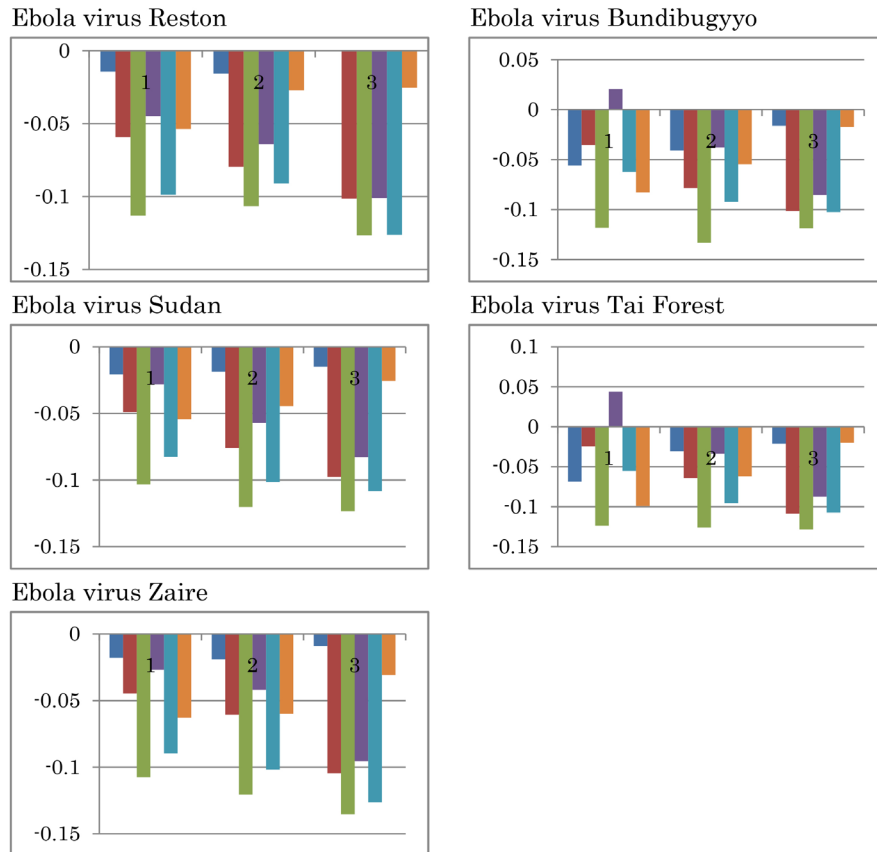
Figure 4. Nucleotide differences in the three fragments of primate mitochondrial genome from 1 to 3. Left to right: (G – C), (G – T), (G – A), (C – T), (C –A) and (T – A).

whose GC content was the highest, both GC and AT inversions occurred. *Mollivirus sibercum*, which had the lowest C content, showed both GC and AT inversions, while DeBrazza's monkey virus 1 showed both GA and GT inversions.

Ebola haemorrhagic fever, caused by the Ebola virus, can be fatal in humans. The Reston, Sudan, and Zaire strains did not show any nucleotide inversion in the three genome fragments (Figure 5). In the Tai Forest and Bundibugyo strains, however, CT inversion was clearly observed in the first fragment, and was accompanied by a decrease in GT content difference. These nucleotide content differences corresponded to the GC contents of the strains: Reston (G: 0.198, C: 0.210) [34], Sudan (G: 0.198, C: 0.216) [35], Zaire (G: 0.198, C: 0.213) [36], Tai Forest (G: 0.192, C: 0.231) [37], and Bundibugyo (G: 0.192, C: 0.228) [37]. An increase in C content and decrease in G content were observed in the Tai Forest and Bundibugyo strains. The calculated GC contents for the strains were: 0.406 (Reston), 0.411 (Zaire), 0.414 (Sudan), 0.420 (Bundibugyo), and 0.423 (Tai Forest). These results may indicate that Ebola virus evolution occurred over a short period of time.

Figure 5. Nucleotide content differences in the three fragments of Ebola virus genomes from 1 to 3. Left to right: (G – C), (G – T), (G – A), (C – T), (C –A) and (T – A).

3.5. Interspecies Evolution

Nucleotide content differences were calculated in vertebrate mitochondria (**Extended Data Figure 7(a)**), invertebrate mitochondria (**Extended Data Figure 7(b)**), non-animal mitochondria (**Extended Data Figure 7(c)**), chloroplasts (**Extended Data Figure 7(d)**) and nuclear DNA (**Extended Data Figure 7(e)**).

3.6. Definitive Universal Equation

Plotting $(X – Y)/(X + Y)$ against $(X – Y)$, the following linear relationship was obtained in mitochondria, chloroplasts, and chromosomes (**Figure 6(a)**), and viruses (**Figure 6(b)**): $(X – Y)/(X + Y) = a (X – Y) + b$, where X and Y are nucleotide contents, and (a) and (b) are constants. As (b) was almost null and (a) was ~2.0, $(X – Y)/(X + Y) ≈ 2.0 (X – Y)$. In these genome analyses, which are independent of Chargaff's parity rules (**Extended Data Figure 8**, left panels), the values of (a) for (G, C), (G, A), (G, T), (C, T), (C, A) and (A, T) were 2.5858, 1.85558, 1.9908, 1.9771, 1.9968 and 1.5689, respectively. Based on these results, $(G + C)$, $(G + A)$, $(G + T)$, $(C + A)$, $(C + T)$ and $(A + T)$ were 0.39, 0.54, 0.50, 0.51, 0.50 and 0.64, respectively. In virus genome analyses (**Extended Data Figure 8**, right panels), the constant values for (a) were 1.9 - 2.1, and the values for $(X + Y)$ were 0.47 - 0.53. In contrast, in the normalization of nucleotide contents $(G + C + A + T = 1)$, as $(G = C)$ and $(A = T)$ based on Chargaff's parity rules, $(2G + 2A = 1)$ is obtained. This equation is altered to $(G + A = 0.5)$. This value is consistent with the value obtained above from genome analyses. Similarly, $(G + T = 0.5)$, $(C + A = 0.5)$ and $(C + T = 0.5)$, although $(G + C)$ and $(A + T)$ cannot be determined. Therefore, the four nucleotide contents are expressed by the following regression lines,
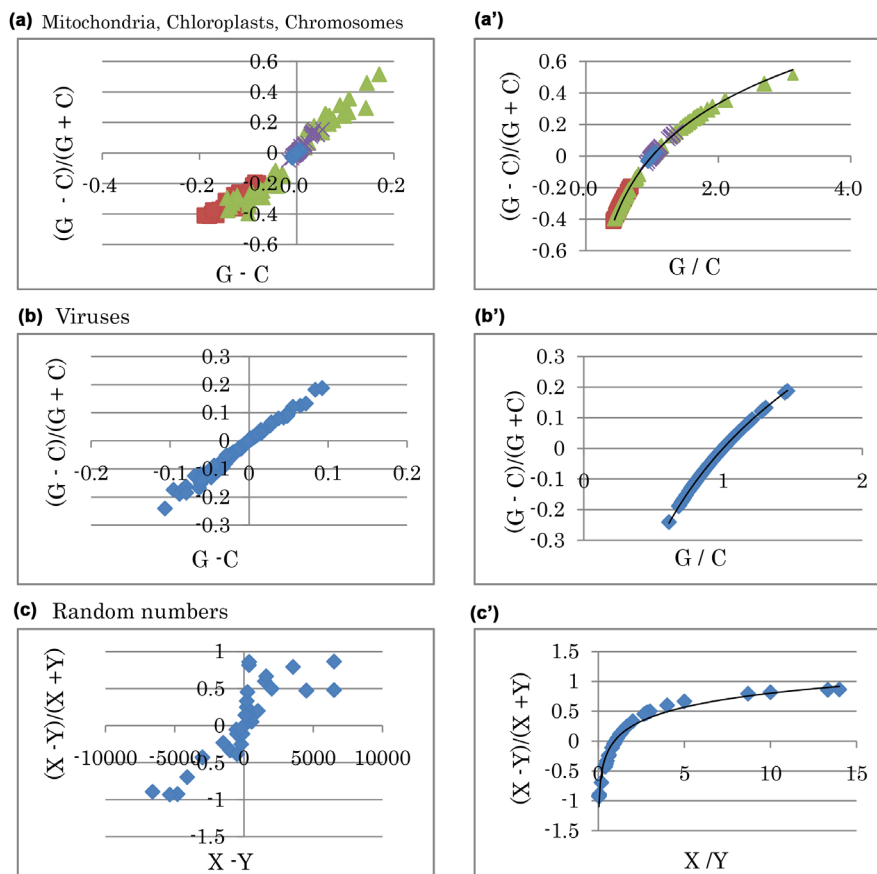
Figure 6. Universal rules. Left side: relationship between (G − C) and (G − C)/(G + C), expressed by a linear regression, and right side: relationship between (G/C) and (G − C)/(G + C), expressed by a logarithmic function. The random numbers were not normalized.

plotted against G content: A = 0.5 − G, T = 0.5 − G, C = G and (G = G). Lines G and C overlap, as do lines A and T, and the former line is symmetrical to the latter against line (y = 0.25). The intercepts of lines G and C are close to the origin, while those of lines A and T are close to 0.5 at the vertical and horizontal axes. All organisms from bacteria to *H. sapiens* are located on the diagonal lines of a 0.5 square, termed the "Diagonal Genome Universe", using the normalized values that obey Chargaff's parity rules [27]. These relationships lead to (G or C) + (A or T) = 0.5. The present results indicate that a linear regression line equation, (X − Y)/(X + Y) = a (X − Y) + b, universally represents all normalised values, including the values deviating from Chargaff's parity rules. This newly discovered equation clearly reflects not only Chargaff's parity rules, based on hydrogen bonding between two nucleotides, but also natural rule.

A linear regression line was not obtained when using randomly chosen value (**Figure 6(c)**). Furthermore, plotting (X − Y)/(X + Y) against (X/Y), the following logarithmic function was obtained for all tested genomes as well as when using randomly chosen values (**Figure 6 (a')-(c')**): (X − Y)/(X + Y) = aln(X/Y) + b. As (b) was almost null and (a) was ~0.5, (X − Y)/(X + Y) ≈ 0.5 ln(X/Y). The ratio between two values, (X/Y), can be expressed by a logarithmic function, ~0.5 ln (X/Y) ≈ (X − Y)/(X + Y). Plotting the GC skew vs. G content, animal mitochondria were classified into two groups: high and low C/G [11]. This fact indicates that the ratio C/G and the GC skew are evolutionarily related to each other. Any change can be expressed universally by a definitive logarithmic function, (X − Y)/(X + Y) = a ln(X/Y) + b. The present results indicate that cellular organelle evolution is strictly controlled under these characteristic rules,
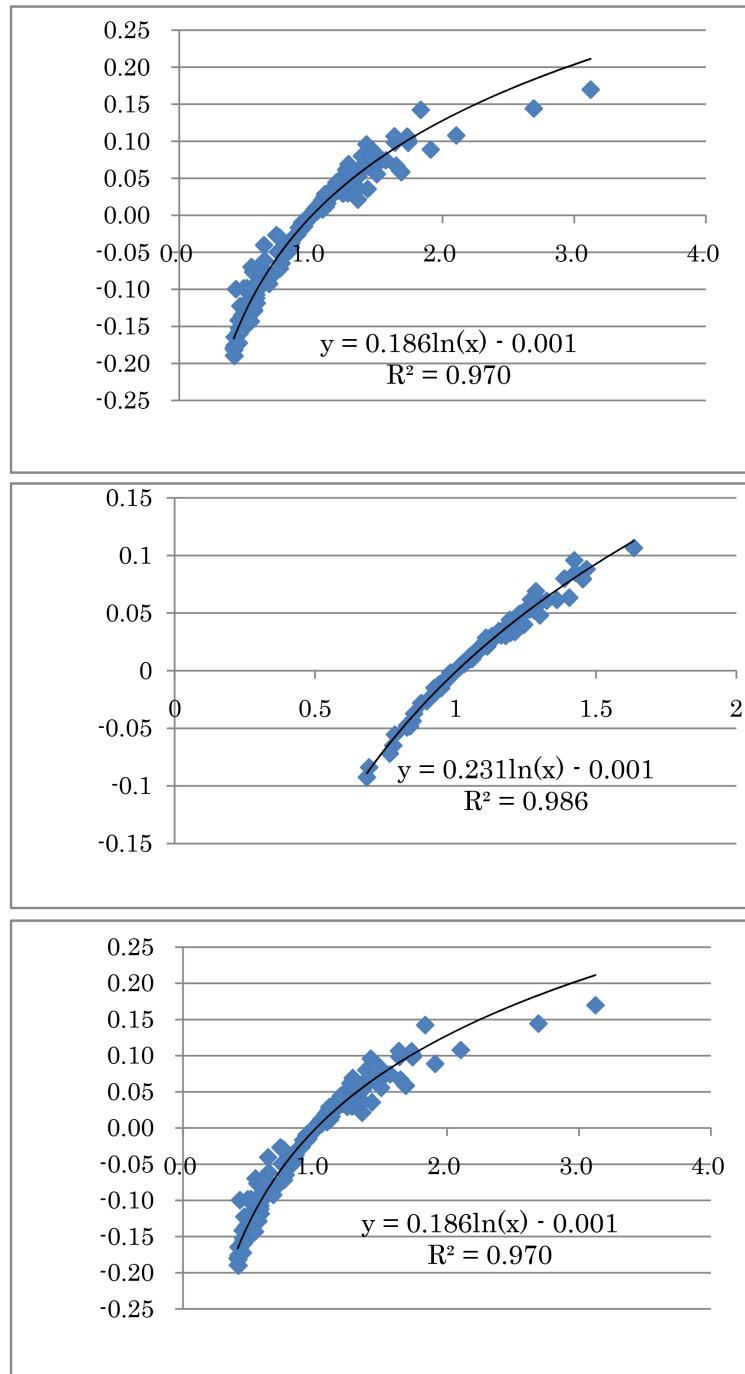
Figure 7. Universal rule governing genome evolution from viruses to *H. sapiens*. Upper pane: cellular organelles, middle panel: viruses, and lower panel: cellular organelles plus viruses.

although non-animal mitochondria, chloroplasts, and chromosomes are controlled under Chargaff's parity rule [17]. The present study clearly shows that biological evolution, which seems to be based on complicated processes, is governed by simple universal equations. In fact, the codon [22] and genome evolution [23] were expressed in complete genomes by linear formulas, and it has been shown that a whole genome

is constructed with gene assembly of small units with a similar amino acid composition [26], as shown in Figure 2 and Extended Data Figure 5.

4. CONCLUSION

Our findings showed that the most primitive extant ancestor of all living organisms is the *M. brevicollis* mitochondrion. In the normalised genome values, the relationship between the nucleotide difference, $(X - Y)$, and $(X - Y)/(X + Y)$, including biological evolution, can be expressed definitively by a linear regression line: $(X - Y)/(X + Y) = a (X - Y) + b$, where $(a \approx 2.0)$ and $(b \approx 0)$ are constants. $(X + Y) =$ approximately 0.5. The nucleotide difference, $(X - Y)$, is generally expressed by a linear regression line that crosses "0" at $(X = Y)$, representing no evolution. In addition, the relationship between the ratio (X/Y) and $(X - Y)/(X + Y)$ can be expressed definitively by a logarithmic function, $(X - Y)/(X + Y) = a \ln(X/Y) + b$, where $(a \approx 0.5)$ and $(b \approx 0)$ are constants. The relationship between the skew "$(X - Y)/(X + Y)$" and the ratio (Y/X) is universally expressed by a logarithmic line that crosses "1" at $(X = Y)$, representing no change. As the left sides, $(X - Y)/(X + Y)$, are equal in both equations, $2.0 (X - Y) \approx 0.5 \ln(X/Y)$. Finally, $(X - Y) \approx 0.25 \ln(X/Y)$ using the normalized values from viruses to *H. sapiens* (Figure 7).

CONFLICTS OF INTEREST

The author declares no conflicts of interest regarding the publication of this paper.

REFERENCES

1. Tsuji, G., Fujii, S., Sunami, T. and Yomo, T. (2016) Sustainable Proliferation of Liposomes Compatible with Inner RNA Replication. *Proceedings of National Academy of Sciences of United States of America*, **113**, 590-595. https://doi.org/10.1073/pnas.1516893113

2. Schopf, J.W. and Barghoorn, E.S. (1967) Alga-Like Fossils from the Early Precambrian of South Africa. *Science*, **3774**, 508-512. https://doi.org/10.1126/science.156.3774.508

3. Nagy, B., Zumberge, J.E. and Nagy, L.A. (1975) Abiotic, Graphitic Microstructures in Micaceous Metaquartzite about 3760 Million Years Old from Southwestern Greenland: Implications for Early Precambrian Microfossils. *Proceedings of National Academy of Sciences of United States of America*, **72**, 1206-1209. https://doi.org/10.1073/pnas.72.3.1206

4. Noffke, N., Christian, D., Wacey, D. and Hazen, R.M. (2013) Microbially Induced Sedimentary Structures Recording an Ancient Ecosystem in the ca. 3.48 Billion-Year-Old Dresser Formation, Pilbara, Western Australia. *Astrobiology*, **13**, 1103-1124. https://doi.org/10.1089/ast.2013.1030

5. Schopf, J.W. (2006) Fossil Evidence of Archaean Life. *Philosophical Transactions of the Royal Society B Biological Sciences*, **361**, 869-885. https://doi.org/10.1098/rstb.2006.1834

6. Gray, M.W. (1992) The Endosymbiont Hypothesis Revisited. *International Review of Cytology*, **141**, 233-357. https://doi.org/10.1016/S0074-7696(08)62068-9

7. Ku, C., Nelson-Sathi, S., Roettger, M., Sousa, F.L., Lockhart, P.J., Bryant, D., *et al.* (2015) Endosymbiotic Origin and Differential Loss of Eukaryotic Genes. *Nature*, **524**, 427-432. https://doi.org/10.1038/nature14963

8. Andersson, S.G., Zomorodipour, A., Andersson, J.O., Sicheritz-Pontén, T., Alsmark, U.C., Podowski, R.M., *et al.* (1998) The Genome Sequence of *Rickettsia prowazekii* and the Origin of Mitochondria. *Nature*, **396**, 133-140. https://doi.org/10.1038/24094

9. Thrash, J.C., Boyd, A., Huggett, M.J., Grote, J., Carini, P., Yoder, R.J., *et al.* (2011) Phylogenomic Evidence for a Common Ancestor of Mitochondria and the SAR11 Clade. *Scientific Reports*, **1**, 13. https://doi.org/10.1038/srep00013

10. Sorimachi, K. and Okayasu, T. (2014) Classification of Non-Animals and Invertebrates Based on Amino Acid Composition of Complete Mitochondrial Genomes. *International Journal of Biology*, **6**, 1-16.

11. Sorimachi, K. (2015) Origine of Life in the Ocean: Direct Derivation of Mitochondria from Primitive Organisms Based on Complete Genomes. *Current Chemical Biology*, **9**, 23-35.
https://doi.org/10.2174/2212796809666150911201738

12. Sorimachi, K. (2010) Genome Data Provides Simple Evidence for a Single Origin of Life. *Natural Science*, **2**, 519-525.

13. Sorimachi, K., Okayasu, T., Ohhira, S., Fukasawa, I. and Masawa, N. (2012) Evidence for the Independent Divergence of Vertebrate and High C/G Ratio Invertebrate Mitochondria from the Same Origin. *Natural Science*, **4**, 479-483.

14. Sorimachi, K. (1999) Evolutionary Changes Reflected by the Cellular Amino Acid Composition. *Amino Acids*, **17**, 207-226. https://doi.org/10.1007/BF01361883

15. Sorimachi, K., Itoh, T., Kawarabaysi, Y., Okayasu, T., Akimoto, K. and Niwa, A. (2001) Conservation of the Basic Pattern of Cellular Amino Acid Composition during Biological Evolution and the Putative Amino Acid Composition of Primitive Life Forms. *Amino Acids*, **21**, 393-399. https://doi.org/10.1007/s007260170004

16. Chargaff, E. (1950) Chemical Specificity of Nucleic Acids and Mechanism of Their Enzymatic Degradation. *Experientia*, **6**, 201-209. https://doi.org/10.1007/BF02173653

17. Rundner, R., Karkas, J.D. and Chargaff, E. (1968) Separation of *B. subtilis* DNA into Complementary Strands. 3. Direct Analysis. *Proceedings of the National Academy of Sciences*, **60**, 921-922.
https://doi.org/10.1073/pnas.60.3.921

18. Sorimachi, K. (2009) A Proposed Solution to the Historic Puzzle of Chargaff's Second Parity Rule. *The Open Genomics Journal*, **2**, 12-14. https://doi.org/10.2174/1875693X00902010012

19. Sorimachi, K., Okayasu, T. and Ohhira, S. (2015) Normalization of Complete Genome Characteristic: Application to Evolution from Primitive Organisms to *Homo sapiens*. *Current Genomics*, **16**, 99-106.
https://doi.org/10.2174/1389202916666150119215716

20. Sorimachi, K. (2014) Indices of Whole Genome Characteristics Based on Amino Acid or Nucleotide Content Ratios: Evolution from Primitive Organisms to *Homo sapiens*. *Journal of Chemical Engineering and Chemistry Research*, **1**, 302-312.

21. Mitchell, D. and Bridge, R. (2006) A Test of Chargaff's Second Rule. *Biochemical and Biophysical Research Communications*, **340**, 90-94. https://doi.org/10.1016/j.bbrc.2005.11.160

22. Sorimachi, K. and Okayasu, T. (2008) Codon Evolution Is Governed by Linear Formulas. *Amino Acids*, **34**, 345-350. https://doi.org/10.1007/s00726-007-0024-3

23. Sorimachi, K. and Okayasu, T. (2008) Universal Rules Governing Genome Evolution Expressed by Linear Formulas. Open *Current Genomics*, **1**, 33-43. https://doi.org/10.2174/1875693X00801010033

24. Nikolaou, C. and Almirantis, Y. (2006) Deviations from Chargaff's Second Parity Rule in Organellar DNA Insights into the Evolution of Organellar Genomes. *Gene*, **381**, 34-41. https://doi.org/10.1016/j.gene.2006.06.010

25. Bell, S.J. and Forsdyke, D.R. (1999) Deviations from Chargaff's Second Parity Rule Correlate with Direction of Transcription. *Journal of Theoretical Biology*, **197**, 63-76.

26. Sorimachi, K. and Okayasu, T. (2003) Gene Assembly Consisting of Small Units with Similar Amino Acid Composition in the *Saccharomyces cerevisia*. *Mycoscience*, **44**, 415-417.
https://doi.org/10.1007/S10267-003-0131-2

27. Sorimachi, K. (2010) Evolution Based on Genome Structure: The "Diagonal Genome Universe". *Natural*

Science, **2**, 1104-1112.

28. Lobry, J.R. (1996) Asymmetric Substitution Patterns in the Two DNA Strands of Bacteria. *Molecular Biology and Evolution*, **13**, 660-665. https://doi.org/10.1093/oxfordjournals.molbev.a025626

29. Tillier, E.R. and Collins, R.A. (2000) The Contributions of Replication Orientation, Gene Direction, and Signal Sequences to Base-Composition Asymmetries in Bacterial Genomes. *Journal of Molecular Evolution*, **50**, 249-257. https://doi.org/10.1007/s002399910029

30. Anderson, S., Bankier, A.T., Barrell, B.G., de Bruijn, M.H., Coulson, A.R., Drouin, J., *et al.* (1981) Sequence and Organization of the Human Mitochondrial Genome. *Nature*, **290**, 457-465. https://doi.org/10.1038/290457a0

31. Fonceca, M.M., Harris, D.J. and Posada, D. (2014) The Inversion of the Control Region in Three Mitogenomes Provides Further Evidence for an Asymmetric Model of Vertebrate mtDNA Replication. *PLoS ONE*, **9**, e106654. https://doi.org/10.1371/journal.pone.0106654

32. Seligmann, H. (2012) Coding Constraints Modulate Chemically Spontaneous Mutational Replication Gradients in Mitochondrial Genomes. *Current Genomics*, **13**, 37-54. https://doi.org/10.2174/138920212799034802

33. King, N., Westbrook, M.J., Young, S.L., Kuo, A., Abedin, M., Chapman, J., *et al.* (2008) The Genome of the Choanoflagellates *Monosiga brevicollis* and the Origin of Metazoans. *Nature*, **451**, 783-788. https://doi.org/10.1038/nature06617

34. Groseth, A., Stroher, U., Theriault, S. and Feldmann, H. (2002) Molecular Characterization of an Isolate from the 1989/90 Epizootic of Ebola Virus Reston among Macaques Imported into the United States. *Virus Research*, **87**, 155-163. https://doi.org/10.1016/S0168-1702(02)00087-4

35. Sanchez, A. and Rollin, P.E. (2005) Complete Genome Sequence of an Ebola Virus (Sudan Species) Responsible for a 2000 Outbreak of Human Disease in Uganda. *Virus Research*, **113**, 16-25. https://doi.org/10.1016/j.virusres.2005.03.028

36. Volchkov, V.E., Volchkova, V.A., Chepurnov, A.A., Blinov, V.M., Dolnik, O., Netesov, S.V., *et al.* (1999) Characterization of the L Gene and 5' Trailer Region of Ebola Virus. *Journal of General Virology*, **80**, 355-362. https://doi.org/10.1099/0022-1317-80-2-355

37. Towner, J.S., Sealy, T.K., Khristova, M.L., Albariño, C.G., Conlan, S., Reeder, S.A., *et al.* (2008) Newly Discovered Ebola Virus Associated with Hemorrhagic Fever Outbreak in Uganda. *PLOS Pathogens*, **4**, e1000212. https://doi.org/10.1371/journal.ppat.1000212

38. Jenkins, P.D., Kilpatorick, C.W., Robinson, M.F. and Timmins, R.J. (2004) Morphological and Molecular Investigations of a New Family, Genus and Species of Rodent (Mammalia: Rodentia: Hystricognatha) from Lao PDR. *Systematics and Biodiversity*, **2**, 419-454. https://doi.org/10.1017/S1477200004001549

39. Dawson, M.R., Marivaux, L., Li, C.K., Beard, K.C. and Métais, G. (2006) Laonastes and the "Lazarus Effect" in Recent Mammals. *Science*, **311**, 1456-1458. https://doi.org/10.1126/science.1124187

40. Gaudin, T.J., Emry, R.J. and Wible, J.R. (2009) The Phylogeny of Living and Extinct Pangolins (Mammalia, Pholidota) and Associated Taxa: A Morphology Based Analysis. *Journal of Mammalian Evolution*, **16**, 235-305. https://doi.org/10.1007/s10914-009-9119-9

41. Gatesy, J., Hayashi, C., Cronin, M.A. and Arctander, P. (1996) Evidence from Milk Casein Genes That Cetaceans Are Close Relatives of Hippopotamid Artiodactyls. *Molecular Biology and Evolution*, **13**, 954-963. https://doi.org/10.1093/oxfordjournals.molbev.a025663

42. Ursing, B.M. and Arnason, U. (1998) Analyses of Mitochondrial Genomes Strongly Support a Hippopotamus-Whale Clade. *Proceedings of the Royal Society B: Biological Sciences*, **265**, 2251-2255. https://doi.org/10.1098/rspb.1998.0567

43. Sorimachi, K. and Okayasu, T. (2013) Phylogenetic Tree Construction Based on Amino Acid Composition and

Nucleotide Content of Complete Vertebrate Mitochondrial Genomes. *IOSR Journal of Pharmacy*, **3**, 51-60.

44. Session, A.M., Uno, Y., Kwon, T., Chapman, J.A., Toyoda, A., Takahashi, S., *et al.* (2016) Genome Evolution in the Allotetraploid Frog *Xenopus laevis*. *Nature*, **538**, 336-343. https://doi.org/10.1038/nature19840

45. Hellsten, U., Simakov, O., Chapman, J., Fahey, B., Gauthier, M.E., Mitros, T., *et al.* (2010) The Genome of the Western Clawed Frog *Xenopus tropicalis*. *Science*, **328**, 633-636. https://doi.org/10.1126/science.1183670

46. Pollet, N. and Mazabraud, A. (2006) Insights from *Xenopus* Genomes. *Genome Dynamics*, **2**, 138-153. https://doi.org/10.1159/000095101

47. Tymowska, J. (1973) Karyotype Analysis of *Xenopus tropicalis* Gray, Pipidae. *Cytogenetics and Cell Genetics*, **12**, 297-304. https://doi.org/10.1159/000130468

48. Sakaguchi, D.S., Murphey, R.K., Hunt, R.K. and Tompkins, R. (1984) The Development of Retinal Ganglion Cells in a Tetraploid Strain of *Xenopus laevis*: A Morphological Study Utilizing Intracellular Dye Injection. *Journal of Comparative Neurology*, **224**, 231-251. https://doi.org/10.1002/cne.902240205

49. Ohno, S. (1970) Evolution by Gene Duplication. Springer, Berlin. https://doi.org/10.1007/978-3-642-86659-3

50. Janvier, P. (2010) Micro RNAs Revive Old Views about Jawless Vertebrate Divergence and Evolution. *Proceedings of the National Academy of Sciences*, **107**, 19137-19138. https://doi.org/10.1073/pnas.1014583107

51. Van Rheede, T., Bastiaans, T., Boone, D.N., Hedges, S.B., de Jong, W.W. and Madsen, O. (2006) The Platypus Is in Its Place: Nuclear Genes and Indels Confirm the Sister Group Relation of Monotremes and Therians. *Molecular Biology and Evolution*, **23**, 587-597. https://doi.org/10.1093/molbev/msj064

EXTENDED DATA TEXT: APPLICATION OF NEWLY DEVELOPED ANALYTICAL METHOD TO VARIOUS ORGANISMS' MITOCHONDRIAL GENOMES

No positive nucleotide content differences were observed in rat (*Rattus norvegicus*) or mouse (*Mus musculus castaneus*) mtDNA. The mitochondria of rodents, hedgehog (*Erinaceus europeaus*), guinea pig (*Cavia porcellus*), and hamster (*Mesocricetus auratus*) showed quite similar nucleotide content difference patterns, except for the presence of AT inversion in the second fragment (Extended Data Figure 2(a)). The Laotian rock rat, *Laonastes aenigmamus*, was described as a new family in 2005 [38], but the classification was changed to Diatomyidae in 2006 [39]. The mtDNA nucleotide content pattern of this rat completely differed from those of the other rodents examined in this study, although it resembled that of the Asian pangolin (*Manis pentadactyla*) (Extended Data Figure 2(a)). Pangolins are classified into two groups, African (*Manis tetradactyla*) and Asian, based on morphological characteristics [40]. In African pangolins, AT content inversion was observed in the second mtDNA fragment, while positive (T − C) differences were observed in all three mtDNA fragments of Asian pangolins. The African pangolin pattern resembled that of prosimian *L. catta*, while the Asian pangolin pattern resembled that of reptile *Heteronotiabinoei* (Bynoe's gecko). Although they belong to the same genus, the two pangolins are remarkably separated in terms of genome biological divergence.

The ocean-dwelling mammal platypus, *Ornithorhynchus anatinus*, and the short-beaked echidna, *Tachyglossus aculeatus*, both belong to the order Monotremata, and seem to be the most primitive mammals. TA inversion occurred in the second mtDNA fragment of both these animals (Extended Data Figure 2(b)). Recently, it was reported that whales are related to hippopotamus [41, 42]. Upon examination in the current study, consistent results were obtained based on Ward's clustering analyses using amino acid compositions or nucleotide contents predicted from complete mitochondrial genomes as traits [43], or using 16S rRNA sequences [43]. However, nucleotide content difference patterns of mtDNA from whale (*Balaenoptera musculus*) and dolphin (*Phocoena phocoena*) species did not resemble that of the pygmy hippopotamus (*Hexaprotodon liberiensis*) or rhinoceros (*Diceros bicornis*) (Extended Data Figure 2(b)). In fact, the whale and dolphin patterns resembled those of primate species *N. coucang* and *T. belangeri*. This discrepancy should be clarified in the future. Nucleotide content difference patterns of mtDNA from birds (*G. gallus* and *T. guttata*) resembled that of primates (Figure 4 and Extended Data Figure 2(b)). This suggests that birds and primates have a similar level of divergence, as described previously [11]. However, a significant difference was observed between the two groups: the (C − T) difference in the second fragment was smaller than that of the first and third fragments in the mtDNA of birds.

Amphibians are classified into the tailed Caudata (*Lyciasalamandra atifi* and *Cynops pyrrhogaster*) and the non-tailed Anua (*Xenopus laevis*, *Rana nigromaculata*, and *Bufo japonicus*) groups. This classification was also confirmed by nucleotide content difference patterns (Extended Data Figure 2(c)). AT, CT, and AC inversions were observed in the latter group, while only AT inversion was observed in the former, except for *X. laevis*. In *X. laevis* and *Dermophis mexicanus* (limbless amphibian) mtDNA, there is no nucleotide content inversion. In contrast, *X. tropicalis* mitochondria have both CT and CA inversion. *Xenopus* species, *X. laevis* [44] and *X. tropicalis* [45], whose complete genome sequences were analysed, are exciting models for the study of gene duplication that can result from hybridization among species [46]. Their mitochondrial nucleotide content difference patterns are quite different from each other. Eventually, the G and C values for *X. tropicalis* (G: 0.144, C: 0.281) mtDNA, which is diploid [47], were higher than those of *X. laevis* (G: 0.135, C: 0.235), which is a pseudo-tetraploid [48]. This fact clearly indicates that the highly evolved *X. tropicalis* diverged via a hybridized ancestor, such as *X. laevis*. Thus, *X. laevis* and *X. tropicalis* seem to be an excellent model to understand evolution based on gene duplication [49].
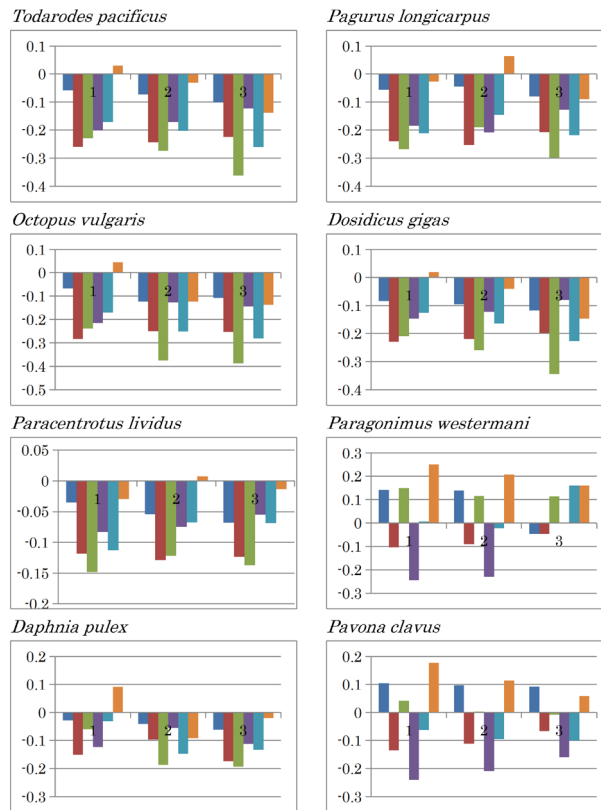
In aquatic vertebrates, the nucleotide content difference pattern of hagfish, *Eptatretus burgeri*, which has been proposed as an ancestor of vertebrates [50], resembled that of lamprey, *Lampetra fluviatilis* (Extended Data Figure 2(e)). Among the Gnathostomata, the mtDNA nucleotide content difference pattern of sea horse (*Hippocampus kuda*) differed from that of seaweed pipefish (*Syngnathus schlegeli*), al-

though they both belong to the class Actinopterygii. The sea horse pattern resembled that of the Chondrichthyes species thorny skate (*Amblyraja radiata*) and brownbanded bambooshark (*Chiloscyllium punctatum*), amphibians (*L. atifi* and *C. pyrrhogaster*), and platypus (*O. anatinus*), which is an ancestor of mammals [51]. The mtDNA nucleotide content difference pattern of the coelacanth *Latimeria chalumnae* resembled that of eel (*Anguilla marmonata*) and lung fish (*Neoceratodus forsteri*), all of which were similar to the mtDNA of reptiles (**Extended Data Figure 2(d)**). These organisms are characterized by AT inversion in the second fragment. These similarities seem to reflect evolutionary processes from aquatic to terrestrial animals [11].

In non-animals, the mtDNA of fungal species *Smittium culisetae* showed the lowest GC and C contents, while the highest values were observed in Cycas species *Cycas taitungensis* mtDNA (**Figure 1**). GC inversion was observed in the former while both GC and AT inversions were observed in the latter (**Extended Data Figure 3**). In chloroplasts, no nucleotide inversion was observed in the red alga *Cyanidioschyzon merolae*, whose GC content was the lowest, while both GC and AT inversions were observed in the green alga *Chara vulgaris*, which had the highest GC content. Regarding C content, however, the *C. vulgaris* chloroplast C content was the lowest amongst the samples examined, while the mtDNA of the green alga *Nephroselmi solivacea* showed the highest C content, and was accompanied by GC and AT inversions.

GC and AT inversions in nucleotide content differences were observed in *U. urealyticum*, *Mycoplasma pulmonis*, *Rickettsia prowazekii*, *Staphylococcus aureus*, and *Escherichia coli*, while minor GC inversion was observed in *S. coelicolor*, which had the highest GC content of all organisms examined (**Extended Data Figure 4**). Amongst the archaea, both GC and AT inversions were observed in *Pyrococcushorikoshii*, while only GC inversion was observed in *Halobacterium*. The inversion patterns based on (G –T), (G – A), (C – T) and (C – A) occurred between low GC content bacteria such as *Ureaplasma urealyticum* (G: 0.131, C:0.127), *Mycoplasma pulmonis* (G: 0.133, C: 0.133), *Rickettsia prowazekii* (G: 0.146, C: 0.144), *Staphylococcus aureus* (G: 0.167, C: 0.165) and *Pyrococcus horikoshii* (G: 0.207, C: 0.212), and high GC content bacteria such as Halobacterium (G: 0.339, C: 0.340)and *Streptomyces coelicor* (G: 0.361, C: 0.359). *Escherichia coli* (G: 0.254, C: 0.254) showed the intermediate pattern.

The microsporidian protozoan *Encephalitozoon cuniculi* is a special eukaryote that lacks mitochondria, although it contains 11 chromosomes. Interestingly, the nucleotide contents of the chromosomes were almost identical (**Extended Data Figure 5**). GC and AT inversions, as well as other nucleotide content inversions, were observed among the 11 chromosomes of *E. cuniculi*, and nucleotide content difference patterns differed among the chromosomes (**Extended Data Figure 5**).

Todarodes pacificus

Pagurus longicarpus

Octopus vulgaris

Dosidicus gigas

Paracentrotus lividus

Paragonimus westermani

Daphnia pulex

Pavona clavus

(a)

Schistosoma mansoni

Acanthaster planci

Geodia neptuni

Haliotis rubra

Tigriopus californicus

Lampsilis ornata

Branchiostoma belcheri

(b)

Saccoglossus kowalevskii

Balnoglossus carnosus

Xenoturbella bocki
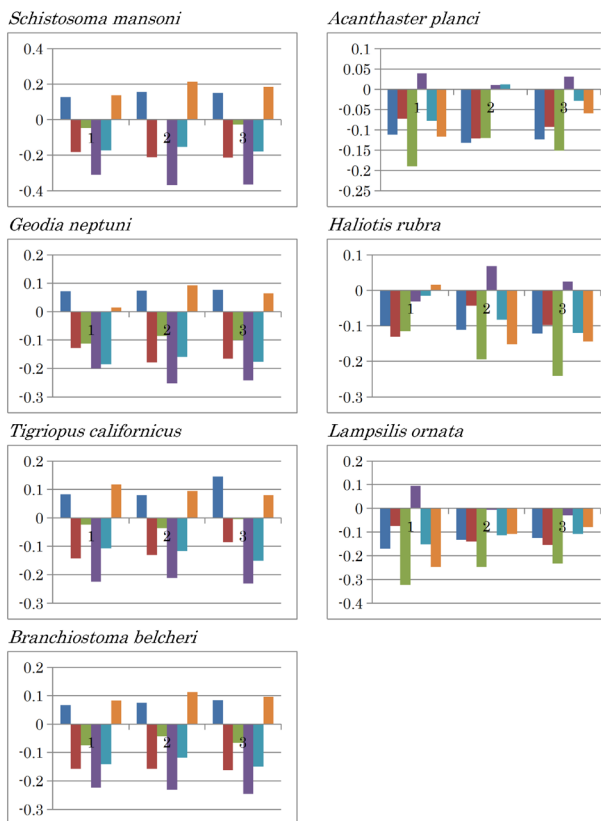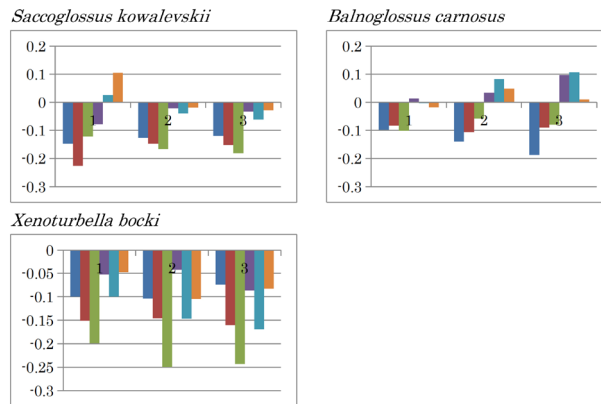
(c)

Extended Data Figure 1. Nucleotide content differences in the three fragments of the mitochondrial genomes of invertebrates. 1, 2, and 3 represent the sequentially divided first, second, and third fragments, respectively. Left to right: (G – C), (G – T), (G – A), (C – T), (C –A), and (T – A).
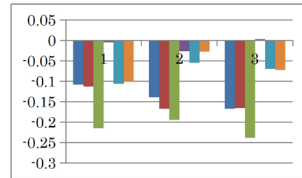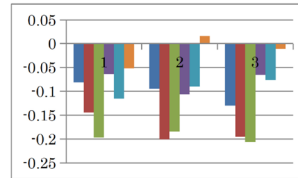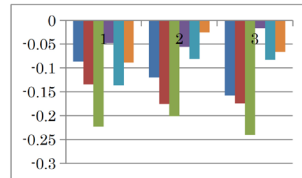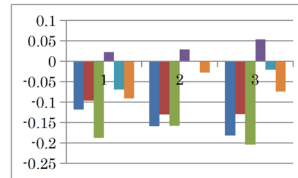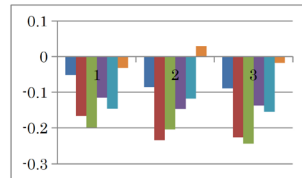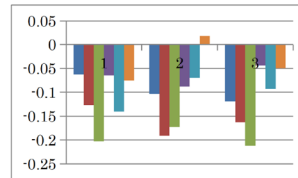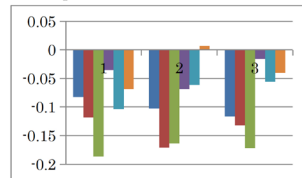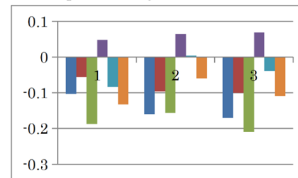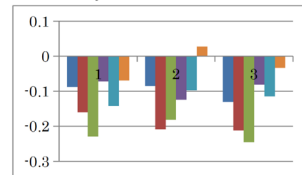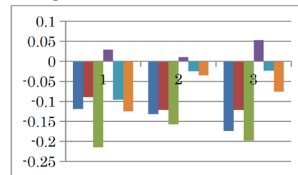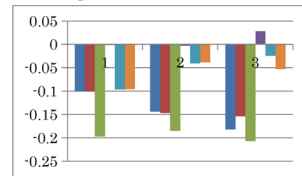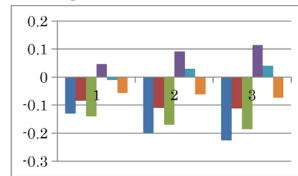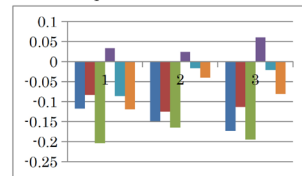
(a)
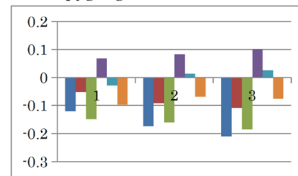
(b)

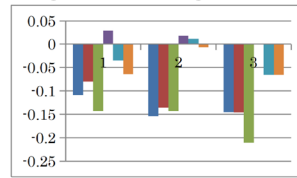Lyciasalamandra atifi
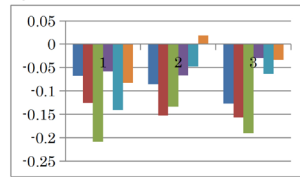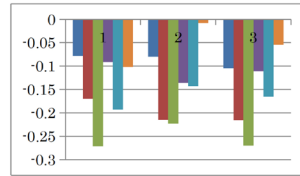
Cynops pyrrhogaster
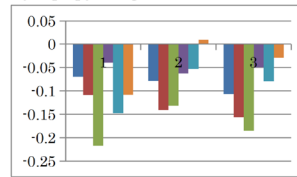
Rana nigromaculata
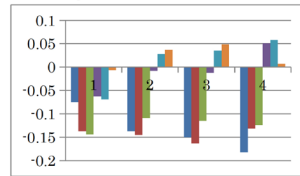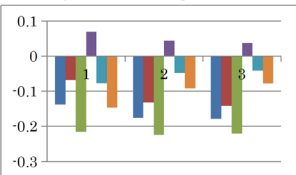
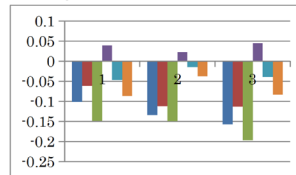Bufo japonicus
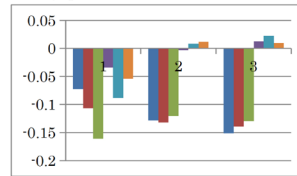
Xenopus laevis
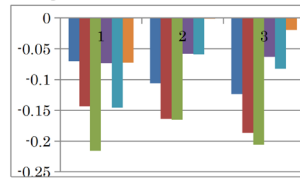
Xenopus (Silurana) tropicalis

Dermophis mexicanus

(c)

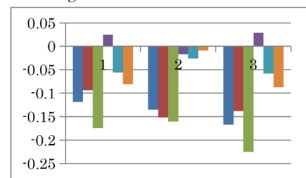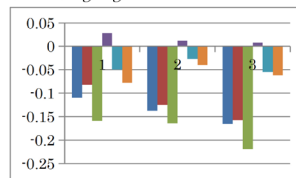Chlamydosaurus kingii

Crocodylus niloticus

Gekko gecko

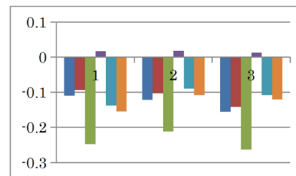Gavialis gangeticus

Heteronotia binoei
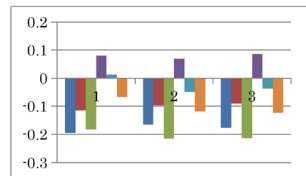
Boa constrictor

Alligator mississippiensis

Chelonia mydas

(d)
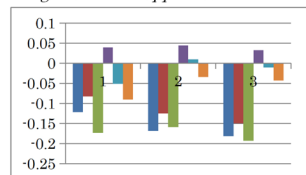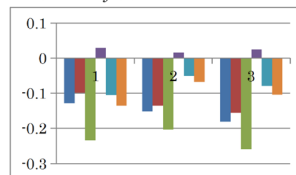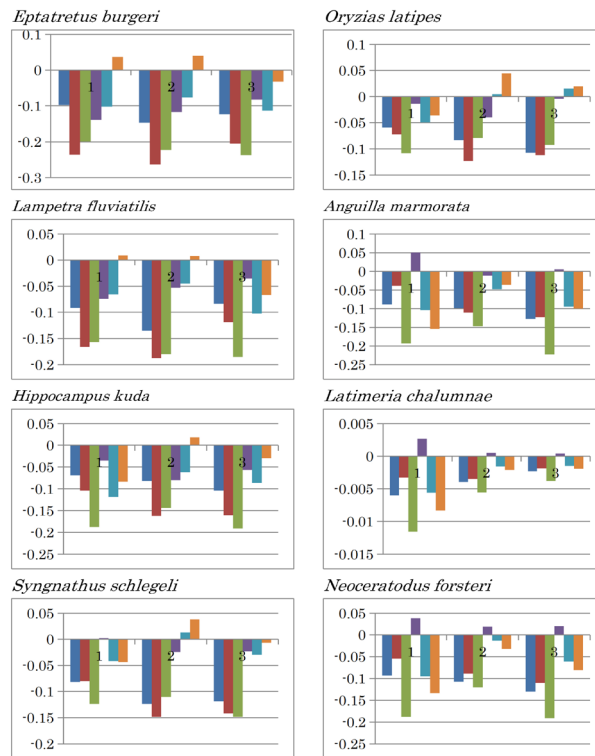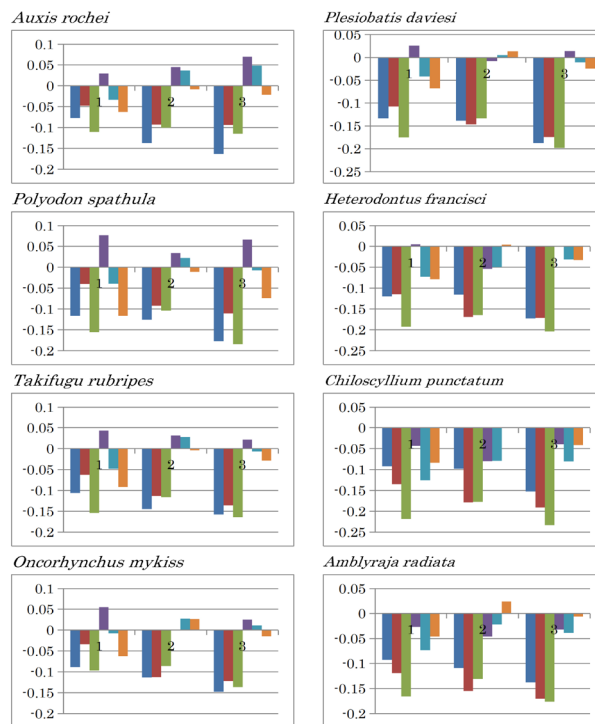
Extended Data Figure 2. Nucleotide content differences in the three fragments of the mitochondrial genomes of vertebrates. 1, 2, and 3 represent the sequentially divided first, second, and third fragments, respectively. Left to right: (G – C), (G – T), (G – A), (C – T), (C –A), and (T – A).
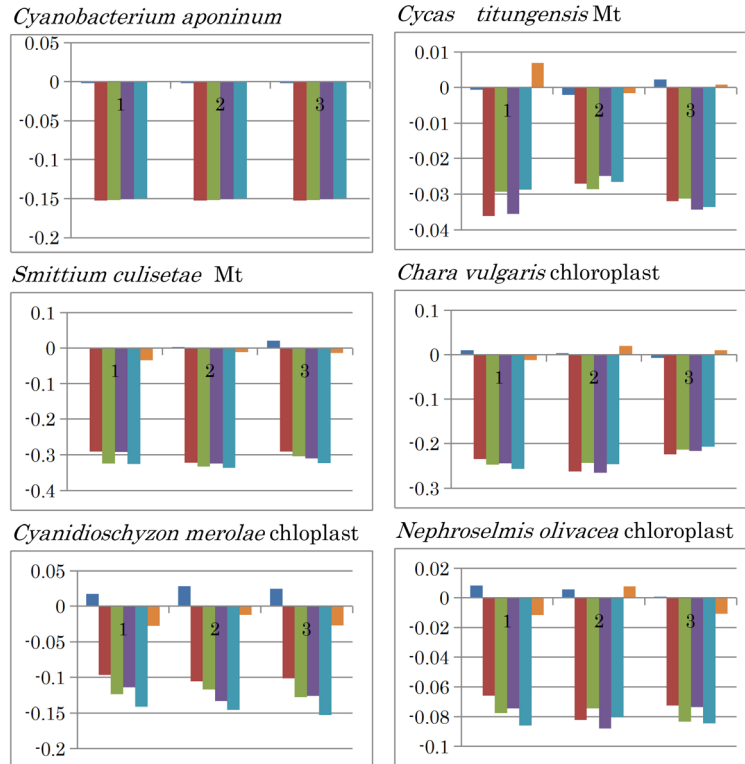
Extended Data Figure 3. Nucleotide content differences in the three fragments of the mitochondrial genomes of non-animal mitochondria and chloroplasts. 1, 2, and 3 represent the sequentially divided first, second, and third fragments, respectively. Left to right: (G – C), (G – T), (G – A), (C – T), (C –A), and (T – A).

Extended Data Figure 4. Nucleotide content differences in the three fragments of bacterial genomes. 1, 2, and 3 represent the sequentially divided first, second, and third fragments, respectively. Left to right: (G – C), (G – T), (G – A), (C – T), (C –A), and (T – A).

Extended Data Figure 5. Nucleotide content differences in the three fragments of the chromosome of microsporidian protozoan species *Encephalitozoo ncuniculi*. Left to right: (G – C), (G – T), (G – A), (C – T), (C –A), and (T – A).
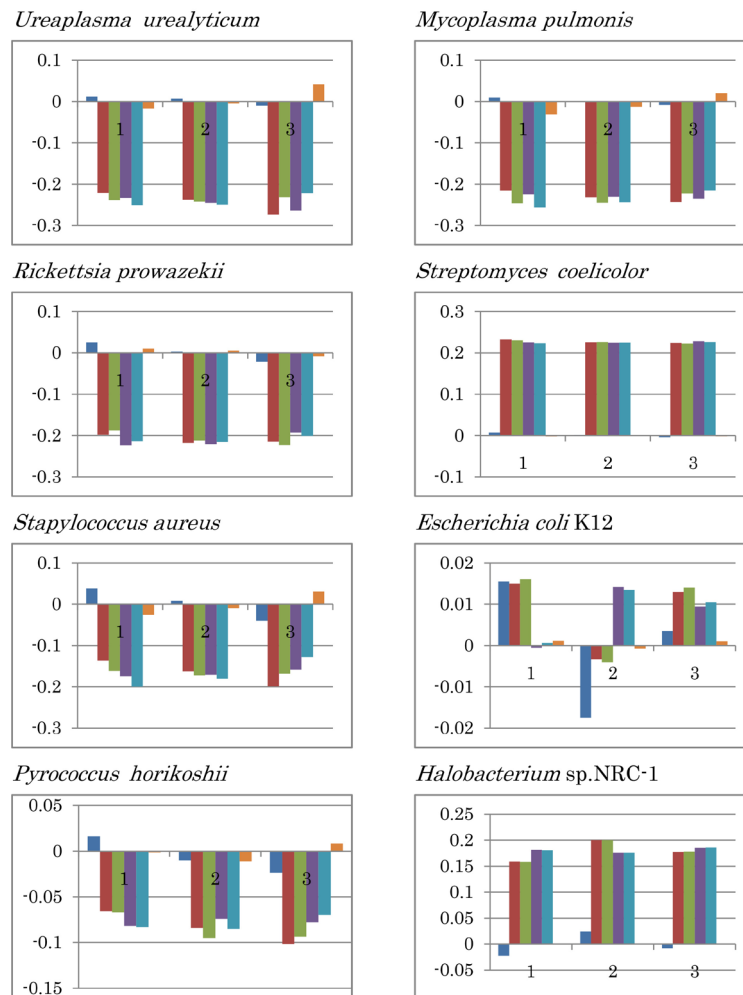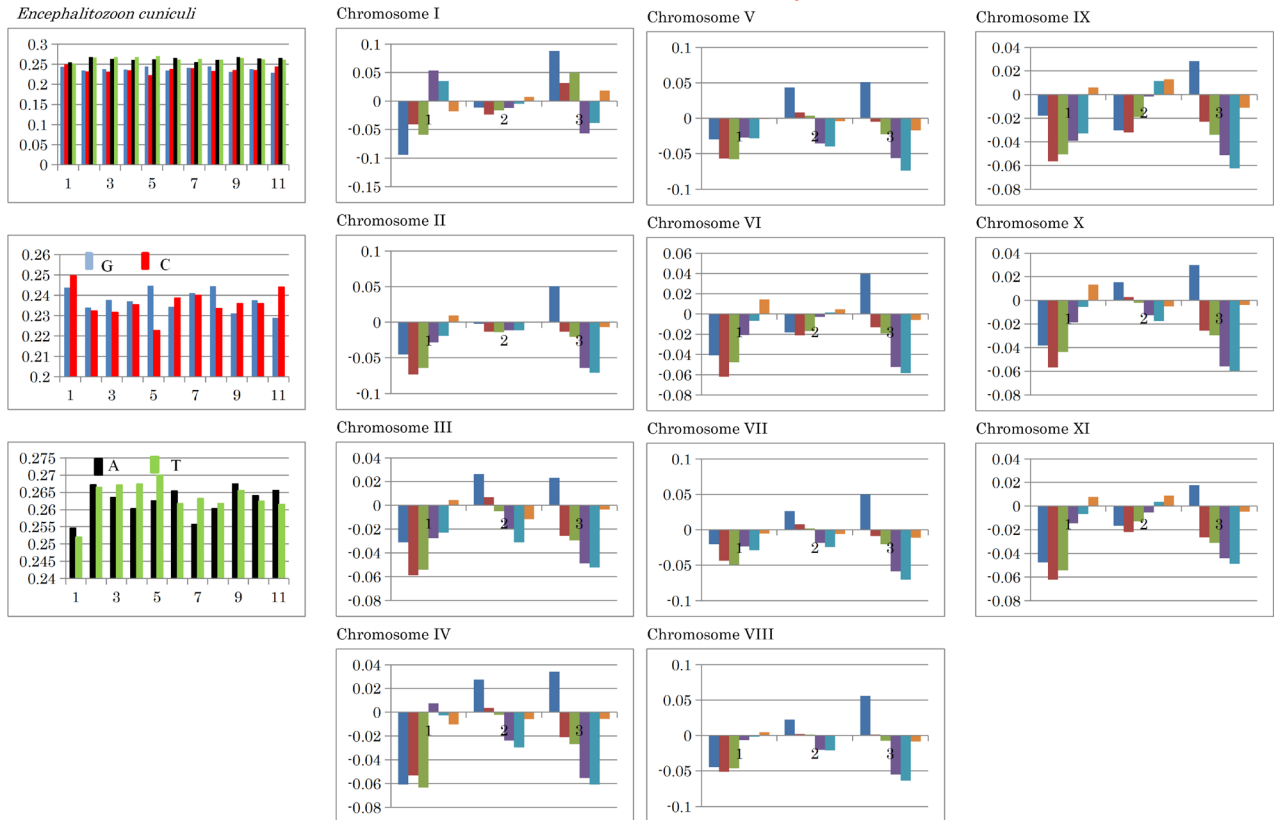

Extended Data Figure 6. Nucleotide content differences in the three fragments of viral genomes. 1, 2, and 3 represent the sequentially divided first, second, and third fragments, respectively. Left to right: (G – C), (G – T), (G – A), (C – T), (C –A), and (T – A).
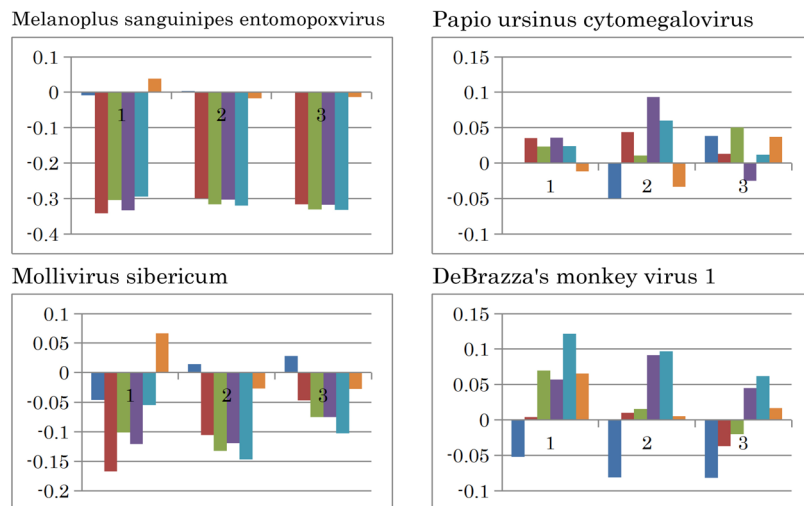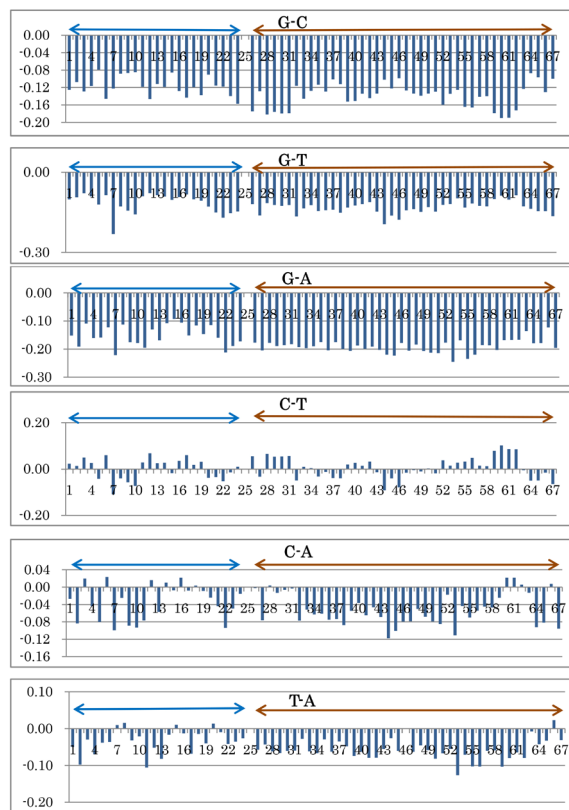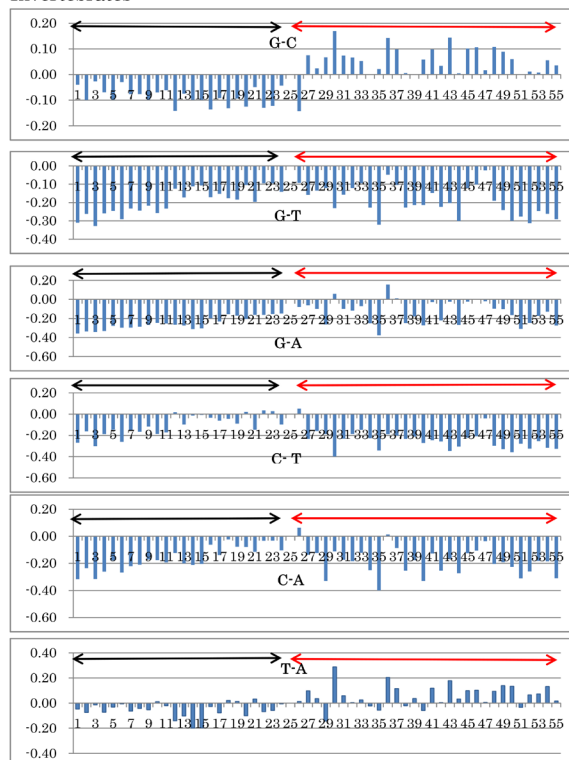
Vertebrates

G-C

G-T

G-A

C-T

C-A

T-A

Vertebrate mitochondria		
No.	Organism	
1	Amia calva	
2	Anguilla marmorata	
3	Auxis rochei	
4	Cyprinus carpio	
5	Danio rerio	
6	Diodon holocanthus	
7	Eptatretus burgeri	
8	Gadus morhua	
9	Hippocampus kuda	
10	Lampetra fluviatilis	
11	Latimeria chalumnae	
12	Myripristis berndti	
13	Neoceratodus forsteri	
14	Oncorhynchus mykiss	
15	Oryzias latipes	
16	Paralichthys olivaceus	
17	Polyodon spathula	
18	Reinhardtius hippoglossoides	
19	Takifugu rubripes	
20	Theragra chalcogramma	
21	Amblyraja radiata	
22	Chiloscyllium punctatum	
23	Heterodontus francisci	
24	Plesiobatis daviesi	
25		
26	Gorilla gorilla	
27	Lemur catta	
28	Homo sapiens	
29	Macaca mulatta	
30	Pan paniscus	
31	Pan troglodytes	
32	Artibeus jamaicensis	
33	Balaenoptera musculus	

34	Bubalus bubalis
35	Canis familiaris
36	Capra hircus
37	Cavia porcellus
38	Cervus nippon yakushimae
39	Diceros bicornis
40	Equus caballus
41	Eumetopias jubatus
42	Hexaprotodon liberiensis
43	Macropus robustus
44	Metachirus nudicaudatus
45	Mus musculus castaneus
46	Ornithorhynchus anatinus
47	Ovis aries
48	Phocoena phocoena
49	Rattus norvegicus
50	Sousa chinensis
51	Sus scrofa
52	Alligator mississippiensis
53	Boa constrictor
54	Crocodylus niloticus
55	Chelonia mydas
56	Chlamydosaurus kingii
57	Gavialis gangeticus
58	Gekko gecko
59	Heteronotia binoei
60	Eudyptes chrysocome
61	Gallus gallus
62	Taeniopygia guttata
63	Bufo japonicus
64	Cynops pyrrhogaster
65	Lyciasalamandra atifi
66	Rana nigromaculata
67	Xenopus laevis

(a)

Invertebrates

G-C

G-T

G-A

C-T

C-A

T-A

Invertebrate mitochondria

No.	Organism		
1	Bombyx mori	29	Dictyostelium discoideum
2	Octopus vulgaris	30	Echinococcus granulosus
3	Drosophila melanogaste	31	Geodia neptuni
4	Metaseiulus occidentalis	32	Lepeophtheirus salmonis
5	Sepioteuthis lessoniana	33	Ligia oceanica
6	Drosophila simulans	34	Malawimonas jakobiformis
7	Bactrocera oleae	35	Monosiga brevicollis
8	Todarodes pacificus	36	Paragonimus westermani
9	Dosidicus gigas	37	Pavona clavus
10	Eriocheir sinensis	38	Physarum polycephalum
11	Pagurus longicarpus	39	Plasmodium vivax
12	Lampsilis ornata	40	Polysphondylium pallidum
13	Tricholepidion gertschi	41	Porites porites
14	Reticulitermes flavipes	42	Reclinomonas americana
15	Reticulitermes hageni	43	Schistosoma mansoni
16	Vargula hilgendorfii	44	Tetrahymena pyriformis
17	Xenoturbella bocki	45	Tigriopus californicus
18	Saccoglossus kowalevskii	46	Tigriopus japonicus
19	Luidia quinalia	47	Trichoplax adhaerens
20	Haliotis rubra	48	Venerupis philippinarum
21	Asterias amurensis	49	Hutchinsoniella macracantha
22	Acanthaster brevispinus	50	Caenorhabditis elegans
23	Acanthaster planci	51	Romanomermis culicivorax
24	Daphnia pulex	52	Strelkovimermis spiculatus
25		53	Thaumamermis cosgrovei
26	Balanoglossus carnosus	54	Trialeurodes vaporariorum
27	Branchiostoma belcheri	55	Varroa destructor
28	Briareum asbestinum		

(b)
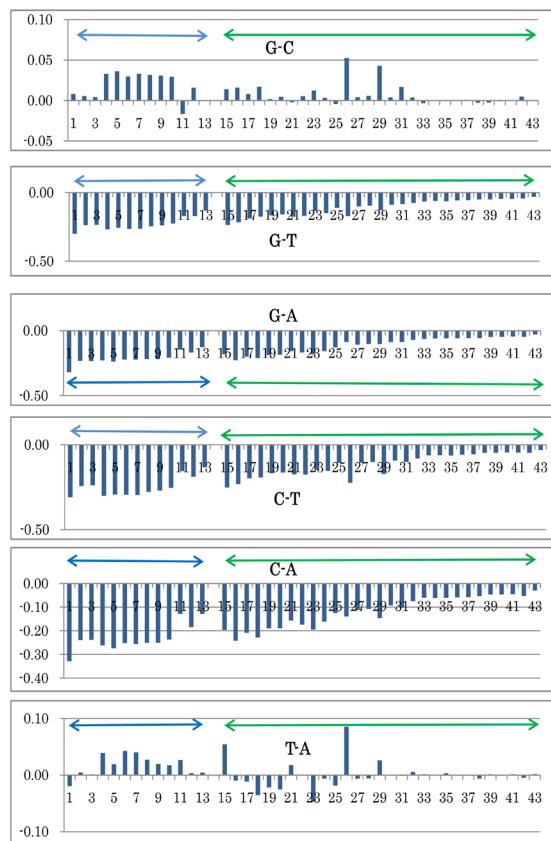
Non-animal Mt

Non-animal mitochondria

No.	Organism		
1	*Smittium culisetae*	23	*Porphyra purpurea*
2	*Rhizopus oryzae*	24	*Chaetosphaeridium globosum*
3	*Pleurotus ostreatus*	25	*Ostreococcus tauri*
4	*Trichophyton rubrum*	26	*Fucus vesiculosus*
5	*Epidermophyton floccosum*	27	*Pseudendoclonium akinetum*
6	*Microsporum canis*	28	*Chlorokybus atmophyticus*
7	*Trichophyton mentagrophytes*	29	*Scenedesmus obliquus*
8	*Sporothrix schenckii*	30	*Chara vulgaris*
9	*Aspergillus fumigatus*	31	*Physcomitrella patens*
10	*Aspergillus niger*	32	*Marchantia polymorpha*
11	*Cryptococcus neoformans*	33	*Sorghum bicolor*
12	*Candida albicans*	34	*Zea mays*
13	*Candida tropicalis*	35	*Oryza sativa*
14		36	*Vitis vinifera*
15	*Emiliania huxleyi*	37	*Triticum aestivum*
16	*Cyanidioschyzon merolae*	38	*Arabidopsis thaliana*
17	*Thalassiosira pseudonana*	39	*Brassica napus*
18	*Rhodomonas salina*	40	*Lotus japonicus*
19	*Mesostigma viride*	41	*Daucus carota*
20	*Phaeodactylum tricornutum*	42	*Chlamydomonas reinhardtii*
21	*Oltmannsiellopsis viridis*	43	*Cycas taitungensis*
22	*Nephroselmis olivacea*		

(c)
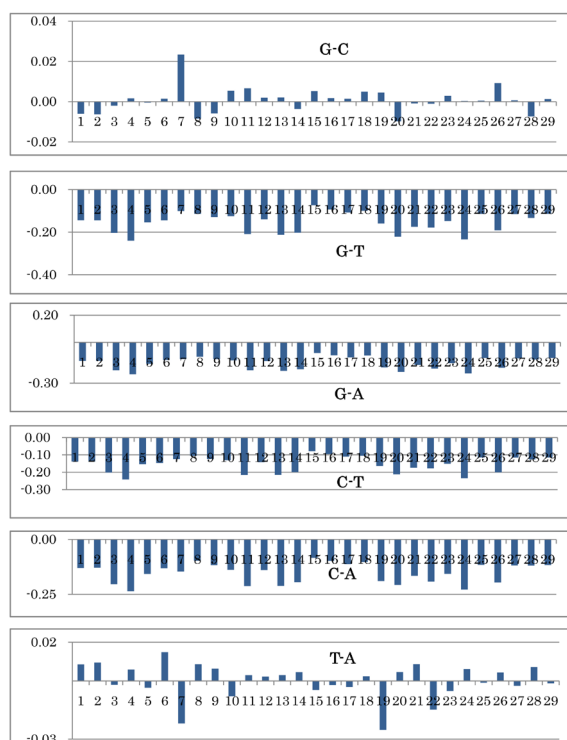
Chloroplasts

No.	Organism		
1	*Arabidopsis thaliana*	15	*Nephroselmis olivacea*
2	*Brassica napus*	16	*Oltmannsiellopsis viridis*
3	*Chaetosphaeridium globosum*	17	*Oryza sativa*
4	*Chara vulgaris*	18	*Ostreococcus tauri*
5	*Chlamydomonas reinhardtii*	19	*Phaeodactylum tricornutum*
6	*Chlorokybus atmophyticus*	20	*Physcomitrella patens*
7	*Cyanidioschyzon merolae*	21	*Porphyra purpurea*
8	*Cycas taitungensis*	22	*Pseudendoclonium akinetum*
9	*Daucus carota*	23	*Rhodomonas salina*
10	*Emiliania huxleyi*	24	*Scenedesmus obliquus*
11	*Fucus vesiculosus*	25	*Sorghum bicolor*
12	*Lotus japonicus*	26	*Thalassiosira pseudonana*
13	*Marchantia polymorpha*	27	*Triticum aestivum*
14	*Mesostigma viride*	28	*Vitis vinifera*
		29	*Zea mays*

(d)

Chlomosomes
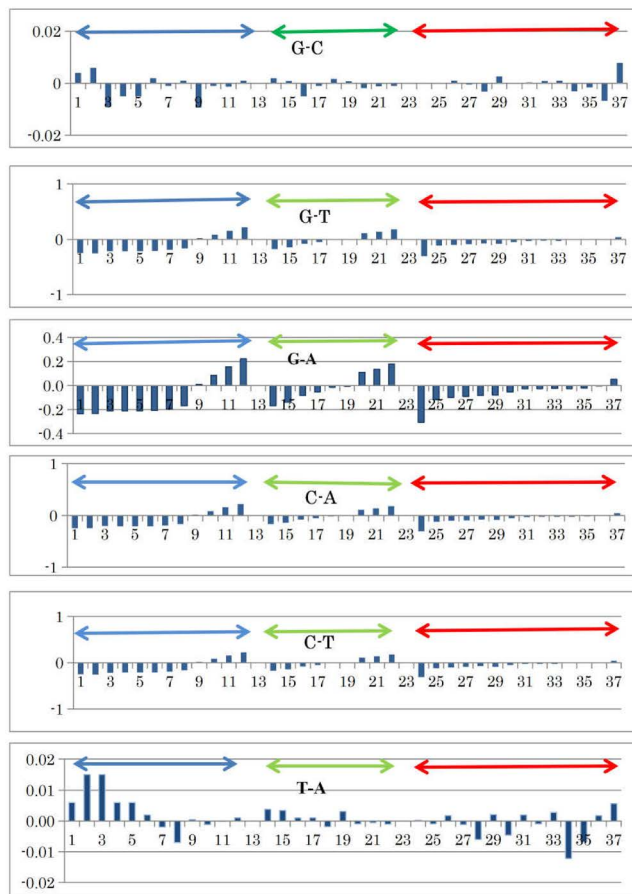
Chlomosomes

No.	Organism
1	*Ureaplasma urealyticum*
2	*Mycoplasma pulmonis*
3	*Clostridium perfringens 1-10*
4	*Rikettsia prowazekii*
5	*Clostridium tetani*
6	*Rickettsia prowazekii*
7	*Campylobacter jejuni*
8	*Staphylococcus aureus*
9	*Escherichia coli K12*
10	*Agrobacterium tumefaciens*
11	*Mycobacterium tuberculosis H37Rv*
12	*Streptomyces coelicolor*
13	
14	*Sulfolobus tokodaii*
15	*Sulfolobus solfataricus P2*
16	*Pyrococcus horikoshii*
17	*Pyrococcus abyssi*
18	*Archaeoglobus fulgidus DSM 4304*
19	*Methanothermobacter*
20	*Methanopyru kandleri AV19*
21	*Methylococcus capsulatus str Bath*
22	*Halobacterium*
23	
24	*Plasmodium falciparum*
25	*Saacharomyces cerevisiae*
26	*Taeniopygia guttata*
27	*Canis lupus familiaris*
28	*Gorilla gorilla*
29	*Ornithorhynchus anatinus*
30	*Equus caballus*
31	*Ovis aries*
32	*Encephalitozoon cuniculi*
33	*Bos taurus*
34	*Homo sapiens*
35	*Mus musculus*
36	*Rattus norvegicus*
37	*Gallus gallus*

(e)

Extended Data Figure 7. Nucleotide differences. (a) Mitochondria of aquatic vertebrates (blue arrow) and terrestrial vertebrates (red arrow). (b) Mitochondria of high C/G invertebrates (black arrow) and low C/G invertebrates (red arrow). (c) Non-animal mitochondria of fungi (blue arrow) and plants (green arrow). (d) Chloroplasts. (e) Chromosomes of prokaryotes (blue arrow), archaea (green arrow), and eukaryotes (red arrow).
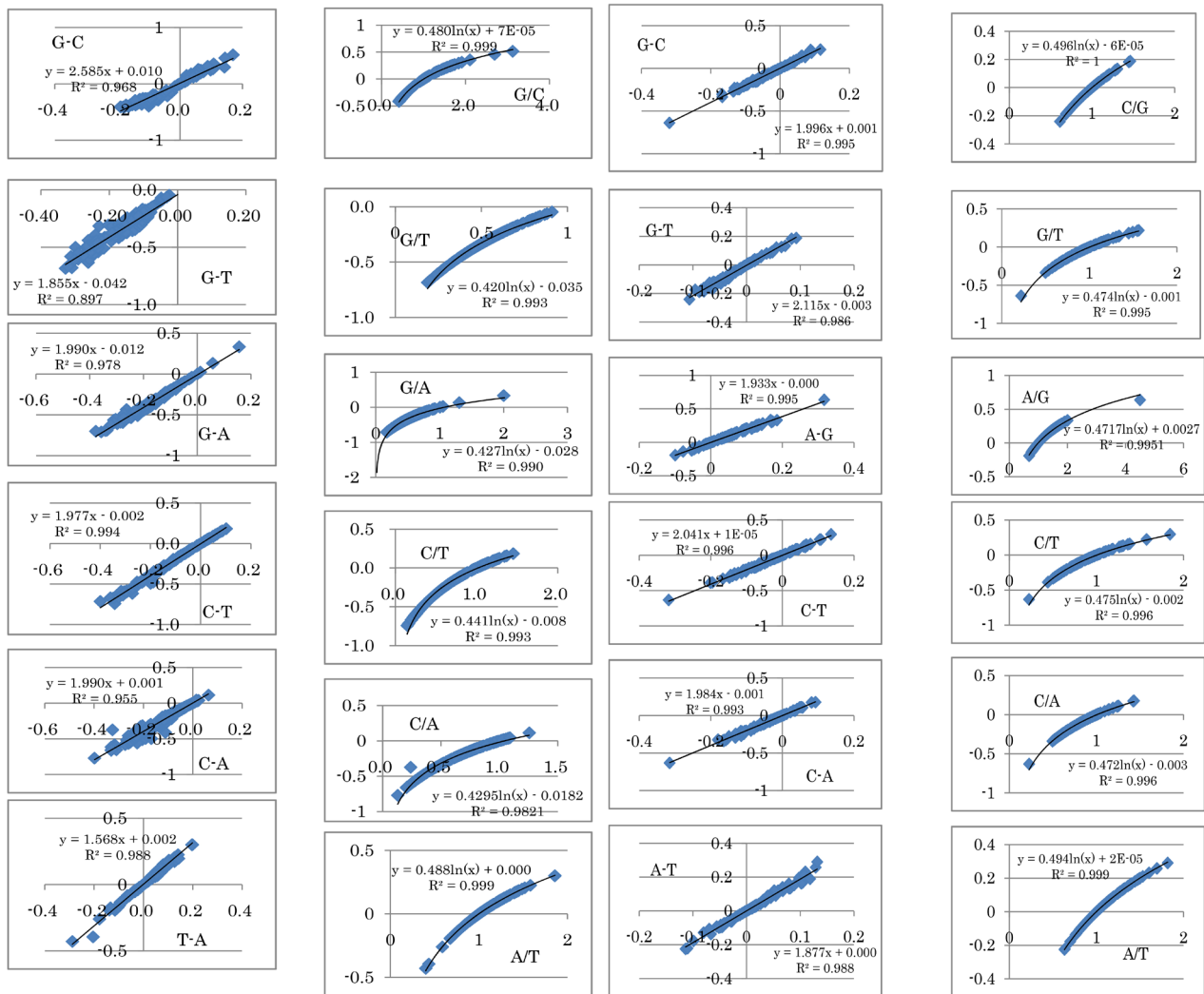
Extended Data Figure 8. Universal rules. Left side: relationship between (X – Y) and (X – Y)/(X + Y), and right side: relationship between (X/Y) and (X – Y)/(X + Y). (a): organelles and (b): viruses.

Extended Data Table 1. List of viruses used.

	Viruses	NC_No.	
1	Gemycircularvirus	NC_027821	sDNA, linear.
2	Pan troglodytes polyomavirus 1	NC_027532	DNA, circular.
3	Mollivirus sibericum isolated P1084-T	NC_027867	DNA, linear.
4	Ruttus norvegicus polyomervirus	NC_027531	DNA, circular
5	Alfalfa leaf curl virus	NC_027714	DNA, circular
6	Poaceae-associated gemycircular virus	NC_027820	DNA, circular
7	New Jersey polyomer-2013 virus	NC_024118	DNA, circular
8	Vaccinia virus	NC_006998	DNA, linear.
9	Taro bacilliform virus	NC_026819	DNA, circular
10	Salmon gill povirus	NC_027707	DNA, linear.
11	Skunk adenovirus PBI virus	NC_027708	DNA, linear.
12	Spodotera exigua nucleopolyhedrovirus	NC_002169	DNA, circular
13	Temperate fruit decay-associated virus	NC_027632	DNA, circular
14	Rubus yellow net virus	NC_026238	DNA, linear.
15	Agrotis segetum nucleopolyhedrovirus	NC_025960	DNA, circular
16	Poecile atricapillus GI trac-associated virus	NC_027776	DNA, circular
17	Rat stool-associated circular ssDNA virus	NC_027777	DNA, circular
18	Scaledrop disease virus	NC_027778	DNA, linear.
19	Human papillomavirus	NC_001693	DNA, linear.
20	Petrochirus diogenes giant hermit crab virus	NC_027780	DNA, circular
21	Aiptasia sp. sea anemone associated virus	NC_027781	DNA, circular
22	Didemnum sp. Sea Squirt associated virus	NC_027782	DNA, circular
23	Littorina sp. Associated circular virus	NC_027783	DNA, circular
24	Callinectes ornatus blue crab virus	NC_027784	DNA, circular
25	Callinectes sapidus virus	NC_027785	DNA, circular
26	Palaemonetes intermedius brackish virus	NC_027786	DNA, circular
27	Gooseberry vein banding virus	NC_018105	DNA, circular
28	Gyrovirus	NC_017091	DNA, circular
29	Tianjin totivirus	NC_017084	dDNA, linear.
30	Canary polyomavirus	NC_017085	DNA, linear.
31	Leucania separata nuclear polyhedrsis virus	NC_008348	DNA, circular.
32	Ageratum leaf curl Cameroon virus	NC_014645	DNA, circular
33	Human papyloma virus type 50 virus	NC_001691	DNA, linear.

Continued

34	Human papylomavirus type 60 virus	NC_001693	DNA, linear.
35	Human papylomavirus type 48 virus	NC_001691	DNA, linear.
36	Citrus chlorotic dwarf associated virus	NC_018151	DNA, circular
37	Sibine fusca densovirus	NC_018399	DNA, linear.
38	Papillo polyxenes densovirus	NC_018450	DNA, linear.
39	French bean severe leaf curl virus	NC_018453	DNA, circular
40	Centrosema yellow spot virus	NC_016998	DNA, circular
41	Cassava mosaic Madagascar A virus	NC_017004	DNA, circular
42	Cassava mosaic Madagascar B virus	NC_017005	DNA, circular
43	Sulfolobus virus	NC_001338	DNA, circular
44	Minute virus of mice	NC_001510	DNA, linear.
45	Bovine parvovirus	NC_001540	DNA, linear.
46	Mastomys natalensis papillomavirus	NC_001605	DNA, circular
47	Simian 40 virus	NC_001669	DNA, circular
48	Rheusus monkey papillomavirus	NC_001678	DNA, circular
49	Adeno-associated virus	NC_001829	DNA, circular
50	Diatraeasaccharalisdensovirus	NC_001899	DNA, circular
51	Melanoplus sanguinipes entomopoxvirus	NC_001993	DNA, linear.
52	Rabbit oral papilloma virus	NC_002232	DNA, linear.
53	Murine pneumotropic virus	NC_001505	DNA, circular
54	Goose pavovirus	NC_001701	DNA, linear.
55	Porcine endogenous retrovirus E	NC_003059	DNA, linear.
56	Dhalia mosaic virus	NC_018616	DNA, circular
57	African pouched rat virus	NC_026439	RNA, circular
58	Equine adenovirus	NC_027705	RNA, linear.
59	Arracacha mottle virus	NC_018176	RNA, linear
60	Phopivirus	NC_027818	RNA, linear.
61	Dengue virus 2	NC_01474	RNA, linear.
62	Parramatta river virus	NC_027817	RNA, linear.
63	West Nile virus	NC_009942	RNA,linear..
64	DeBrazza's monkey virus	NC_026509	RNA, linear.
65	Fizroy River virus	NC_027700	RNA, linear.
66	Trichomonas vaginalis virus	NC_027701	RNA, linear.
67	Luffa aphid-borne yellows virus	NC_027703	RNA, linear.

Continued

68	Lettuce italian necrotic virus	NC_027706	RNA, linear.
69	Spanish goat encephalitis virus	NC_027709	RNA, linear.
70	Dromedary astrovirus	NC_027711	RNA, linear.
71	Tobaco virus	NC_027712	RNA, linear.
72	Beauveria bassiana RNA virus	NC_027428	RNA, linear.
73	Fe-de-lance virus	NC_005084	RNA, linear.
74	Persimmon crytic virus	NC_017988	RNA, linear.
75	Rhizoctonia solani RNA virus	NC_027529	RNA, linear.
76	Japanese holly fern mottle virus	NC_013134	RNA, linear.
77	Chaoyang virus	NC_017086	RNA, linear.
78	Rabit coronavirus	NC_017083	RNA, linear.
79	SFTS virus seg, L	NC_018136	RNA, linear.
80	SFTS virus seg, S	NC_018137	RNA, linear.
81	SFTS virus, seg, M	NC_018138	RNA, linear.
82	Chrysantheum B virus	NC_009087	RNA, linear.
83	Marine birna virus	NC_008019	RNA, linear.
84	Viral hemorrhagic septicemia virus	NC_000855	RNA, linear.
85	Saccharomyces cerrevisiae virus	NC_001641	RNA, linear.
86	Bat hepevirus	NC_018382	RNA, linear.
87	Grapevine Anatolian ringspot RNA 1virus	NC_018383	RNA, linear.
88	Sweet potato virus C-6	NC_018448	RNA, linear.
89	Grapevine virus F	NC_018458	RNA,.linear.
90	Porcine coronavirus HKU 15	NC_016990	RNA, linear.
91	White-eye coronavirus	NC_016991	RNA, linear.
92	Magpie-robin coronvirus	NC_016993	RNA, linear.
93	Night-heron coronavirus	NC_016994	RNA, linear.
94	Wigeon coronavirus	NC_016995	RNA, linear.
95	Common-moorhen coronavirus	NC_016996	RNA, linear.
96	Donggang virus	NC_016997	RNA, linear.
97	Clitocybe odora virus	NC_017003	RNA, linear.
98	Mercadeo virus	NC_027819	RNA, linear.
99	Opium poppy mosaic virus	NC_027710	RNA, linear.
100	Groundnut rossete virus	NC_003603	RNA, linear.