1

2

3

4 **The most primitive extant ancestor of organisms and**
5 **discovery of definitive evolutionary equations**

6

7

8

9 Kenji Sorimachi

10

11

12

13 Educational Support Center, Dokkyo Medical University, Mibu, Tochigi 321-0293,

14 Japan,

15 and Bioresearch Laboratory, Gunma Agriculture and Forest Development, Co. Ltd,

16 Takasaki, Gunma 370-0851, Japan

17

18 Short title: Development of new analytical method

19 Abstract

20 Organisms are classified into three domains, Prokaryota, Archaea, and
21 Eukaryota, and their evolutionary divergence has been characterized based on
22 morphological and molecular features using rationale based on Darwin's theory of
23 natural selection. However, universal rules that govern genome evolution have not been
24 identified. Here, a simple, innovative approach has been developed to evaluate
25 biological evolution initiating the origin of life: whole genomes were divided into
26 several fragments, and then differences in normalized nucleotide content between
27 nucleotide pairs were compared. Based on nucleotide content structures, *Monosiga*
28 *brevicollis* mitochondria may be the most primitive extant ancestor of the species
29 examined here. The two normalized nucleotide contents are universally expressed by a
30 linear regression line, $(X - Y)/(X + Y) = a(X - Y) + b$, where X and Y are nucleotide
31 contents and (a) and (b) are constants. The value of (G + C), (G + A), (G + T), (C + A),
32 (C + T) and (A + T) was ~0.5. Plotting $(X - Y)/(X + Y)$ against X/Y showed a
33 logarithmic function $(X - Y)/(X + Y) = a \ln X/Y + b$, where (a) and (b) are constant.
34 Nucleotide content changes are expressed by a definitive equation, $(X - Y) \approx 0.25$
35 $\ln(X/Y)$.

36

37 Introduction

38 Primitive organisms might appear after long periods of chemical evolution,
39 during which various organic compounds were accumulated. Recently, it was reported
40 that vesicles consisting of lipids and polynucleotides spontaneously replicated under
41 experimental conditions [1]. These pseudo-organisms were the first living things to
42 appear on Earth; however, we cannot trace their origins because the original assembly
43 of chemical compounds was not stable. Only those primitive organisms that established
44 a method of self-replication could survive and continue to evolve. Therefore, it is
45 difficult to identify the true primitive ancestors of life on Earth, as their very make-up
46 would probably be unstable in the current environmental conditions. Although fossilized
47 traces of very early organisms have been found in sedimentary rocks dating from
48 3.1–3.7 billion years ago [2-5], they do not appear to be the origin of life. It is likely that
49 only primitive organisms that had cell walls could leave a fossil record, which may rule
50 out finding any traces of simpler organisms.

51 Several key eukaryotic organelles originated from symbioses between separate
52 single-celled organisms [6,7]. For example, mitochondria developed from the
53 proteobacterium *Rickettsia* or its relatives [8,9]. The *Reclinomonas americana* (Protist)
54 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]. 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

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

88

89 **Materials and methods**

90 All genome sequences were obtained from the National Center for

91 Biotechnology Information GenBank database (<http://www.genome.jp/ja>). The
 92 nucleotide contents of all genomes, which are listed in Extended Data Figs, were
 93 normalized ($G + C + A + T = 1$) because normalized values are independent of species
 94 and genome sizes [19,20]. A whole genome was divided into several fragments to
 95 evaluate intra-genome alterations due to evolution, and the differences between the two
 96 nucleotide contents of six pairs (($G - C$), ($G - T$), ($G - A$), ($C - T$), ($C - A$), ($T - A$))
 97 were calculated for all fragments. In addition, mitochondria were divided into three
 98 equal-size fragments. To easily understand evolution at a glance, these six nucleotide
 99 differences were graphically expressed with a single pattern. This appears to be a novel
 100 approach. Calculations were carried out with a personal computer, TOSHIBA dynabook
 101 T552 (Tokyo, Japan), Windows 10 installed.

102

103 **Results and discussion**

104 **Primitive ancestor**

105 Chargaff's parity rules, where ($G = C$) and ($T = A$), were the original rules for
 106 inter- [16] and intra- [17] molecular relationships of DNA, and are applicable to
 107 chromosomal DNA [21-23]. The normalized nucleotide contents (C , T and A), predicted
 108 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] (Fig 1a). Vertebrate mitochondria overlapped with the high C/G invertebrates, which have high C content (Fig 1a), 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 (Fig 1a). This is consistent with our previous findings [11,12]. In addition, we found that nucleotide content relationships in viral DNA were heteroskedastic (Fig 1b).

121

Fig. 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,

127 chloroplasts (diamond), and chromosomes (square). Red square represents *Monosiga*
 128 *brevicollis* mitochondrion. Right upper panel (a): C content plotted against G content in
 129 viruses. Middle panels (c and d): Horizontal and vertical axes represent (G + C) and
 130 (A + T) contents, respectively. Lower panels (e and f): Horizontal and vertical axes
 131 represent (C) and (G + A + T) contents, respectively.

132

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

145 highest (Fig 1d).

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

155

156 **Genome evolution**

157 The complete genome is represented by four nucleotide contents based on more
158 than a certain amount of randomly chosen fragments, as well as on completely linear
159 fragments [26,27]. Therefore, when the whole genome of bacterium *Ureaplasma*
160 *urealyticum* (G; 0.131, C; 0.127) was sequentially divided into nine equal fragments, the
161 amounts of the four nucleotides in each fragment were quite similar (Fig 2a). This is
162 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 fragments 5 and 6 of the *U. urealyticum* genome (Fig 2b and 2c). The ratios of $(G - C)/(G + C)$ and $(A - T)/(A + T)$ 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 strands²⁷. 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/G¹¹. In *M. brevicollis* mitochondria, the nine DNA fragments showed almost the same nucleotide contents (Fig. 2d), as was observed in *U. urealyticum* (Fig 2a). However, GC and AT content difference inversions were not observed in *M. brevicollis* mtDNA (G: 0.081, C: 0.059) (Fig 2e and 2f). 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.

181

182 **Fig. 2. Nucleotide content was sequentially divided into nine fragments.** Left side:

183 *Ureaplasma urealyticum* mitochondrion, right side: *Monosiga brevicollis* mitochondrion.

184 G: blue, C: red, A: black, and T: green. Vertical axis represents nucleotide content.

185

186 **Organelle evolution**

187 Complete mitochondrial genomes were investigated (Fig 3, left panel). To
 188 allow simple visual comparison of inter- and intra-species genome structures, genomes
 189 were sequential divided into three fragments throughout subsequent analyses, from
 190 which three separate patterns emerged. Six nucleotide content differences were
 191 observed among the mitochondria of the four species (*M. brevicollis*, *P. pallidum*, *D.*
 192 *discoideum* and *R. americana*) (Fig 3, right panels). The six nucleotide content patterns
 193 were conserved within the three fragments among the four species. No inversion of
 194 nucleotide content differences was observed in the mtDNA of *M. brevicollis* (G: 0.081,
 195 C: 0.059), the mycetozoan *Polysphondylium pallidum* (G: 0.143, C: 0.085), or
 196 *Dictyostelium discoideum* (G: 0.171, C: 0.104) (Fig 3), although differences in (G – C)
 197 and (T – A) values for *M. brevicollis* mtDNA were the lowest amongst these species.
 198 Based on genome sequencing, choanoflagellates are most closely related to animals [33].

199 As the nucleotide content difference patterns of the three fragments were almost
200 identical for these three species, their nucleotide distributions were judged to be
201 homogeneous, indicating nucleotide content symmetry. Thus, these mitochondria are
202 likely to be primitive. Consistent results were obtained from Ward's clustering analysis
203 using amino acid compositions predicted from complete mitochondrial genomes as
204 traits¹¹. These findings indicate that the *M. brevicollis* mitochondrion is the most
205 primitive among the three. In contrast, AT inversion was observed in the third fragment
206 of *Reclinomonas americana* mtDNA (G: 0.148, C: 0.114), which has previously been
207 proposed as a mitochondrial ancestor [8]. However, differences in (G – C) and (T – A)
208 values in *R. americana* mtDNA were smaller than those in the mtDNA of the previous
209 three organisms. Nucleotide content inversion causes significant differences in
210 nucleotide content patterns as a result of unsymmetrical nucleotide content. Thus, the *R.*
211 *americana* mitochondrion is probably more evolved than the former three mitochondria.
212 In addition, AT inversion occurred in the following more highly evolved organisms:
213 Mollusca species, squid (*Todarodes pacificus*), octopus (*Octopus vulgaris*),
214 Echinodermata species, sea urchin (*Paracentrotus lividus*), water flea (*Daphnia pulex*),
215 hermit crab (*Pagurus longicarpus*), and Humboldt squid (*Dosidicus gigas*) (S1 Fig). In
216 addition, large positive (G – A) values in the three fragments were observed in

217 *Paragonimus westermani*, while large positive (G – C) and (A – T) values in the three
218 fragments were observed for the mtDNA of representatives of the following phyla:
219 Cnidaria (*Pavona clavus*), Platyhelminthes (*Schistosoma mansoni*), Porifera (*Geodia*
220 *neptuni*), Arthropoda (*Tigriopus californicus*), and Chordata (*Branchiostoma belcheri*)
221 (S1 Fig.).

222

223 **Fig. 3. Nucleotide content differences in complete mitochondrial genomes (left side)**
224 **and the three sequential fragments of each mitochondrial genome (right side).** Left
225 to right: (G – C), (G – T), (G – A), (C – T), (C – A) and (T – A).

226

227 A positive (C – T) value was characteristically observed in the three fragments
228 of Echinodermata species *Acanthaster planci*, the second and third fragments of
229 Mollusca species *Haliotis rubra*, and in the first fragment of Mollusca species *Lampsilis*
230 *ornata*. CT inversion occurred in *H. rubra* and *L. ornata* mtDNA (S1 Fig.). The
231 nucleotide content difference patterns of the mtDNA of hemichordates *Saccoglossus*
232 *kowalevskii* and *Balanoglossus carnosus* differed from each other. Both AT and CT
233 inversions occurred in the first mtDNA fragment of *S. kowalevskii*, while large positive
234 (C – T) and (C – A) value differences occurred in the second and third fragments. AC

235 and AT inversions were observed in *B. carnosus* mtDNA (S1 Fig.). Neither nucleotide
236 inversion nor positive nucleotide differences were observed in the mtDNA of
237 deuterostom, *Xenoturbella bocki*.

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

251

252 **Fig. 4. Nucleotide differences in the three fragments of each mitochondrial genome.**

253 Left to right: (G – C), (G – T), (G – A), (C – T), (C – A) and (T – A).

254

255 The mitochondria of other vertebrates (S1 Text.), rodents (S2 Fig),
 256 ocean-dwelling mammals, cetaceans, aves (S2 Fig.), amphibians (S2 Fig.), reptiles (S2
 257 Fig.), and fishes (S2 Fig.) were also examined. In these organelles, differences in (G –
 258 C) and (A – T) or in other nucleotide inversions, as well as GC and AT inversions,
 259 increased along with evolution. Consistent results were also obtained from non-animal
 260 mitochondria and chloroplasts (S3 Fig.), as well as prokaryotes (S4 Fig.). In the
 261 microsporidian protozoan *Encephalitozoon cuniculi*, GC and AT inversions, as well as
 262 other nucleotide content inversions, were observed (S5 Fig.).

263

264 **Virus evolution**

265 The *M. sanguinipes* Entomopoxvirus genome had the lowest G content among
 266 the viruses examined (Fig. 1), and AT inversion was also observed (S6 Fig.). In *P.*
 267 *ursinus* Cytomegalovirus, whose GC content was the highest, both GC and AT
 268 inversions occurred. *Mollivirus sibiricum*, which had the lowest C content, showed both
 269 GC and AT inversions, while DeBrazza's monkey virus 1 showed both GA and GT
 270 inversions.

271 Ebola haemorrhagic fever, caused by the Ebola virus, can be fatal in humans.

272 The Reston, Sudan, and Zaire strains did not show any nucleotide inversion in the three

273 genome fragments (Fig. 5). In the Tai Forest and Bundibugyo strains, however, CT

274 inversion was clearly observed in the first fragment, and was accompanied by a

275 decrease in GT content difference. These nucleotide content differences corresponded to

276 the GC contents of the strains: Reston (G: 0.198, C: 0.210)[34], Sudan (G: 0.198, C:

277 0.216)[35], Zaire (G: 0.198, C: 0.213)[36], Tai Forest (G: 0.192, C: 0.231)[37], and

278 Bundibugyo (G: 0.192, C: 0.228)[37]. An increase in C content and decrease in G

279 content were observed in the Tai Forest and Bundibugyo strains. The calculated GC

280 contents for the strains were: 0.406 (Reston), 0.411 (Zaire), 0.414 (Sudan), 0.420

281 (Bundibugyo), and 0.423 (Tai Forest). These results may indicate that Ebola virus

282 evolution occurred over a short period of time.

283

284 **Fig. 5. Nucleotide content differences in the three fragments of Ebola virus**

285 **genomes.** Left to right: (G – C), (G – T), (G – A), (C – T), (C – A) and (T – A).

286

287 Interspecies evolution

288 Nucleotide content differences were calculated in vertebrate mitochondria (S7

289 Fig.), invertebrate mitochondria (S7 Fig.), non-animal mitochondria (S7 Fig.),
290 chloroplasts (S7 Fig.) and nuclear DNA (S7 Fig.).

291

292 **Definitive universal equation**

293 Plotting $(X - Y)/(X + Y)$ against $(X - Y)$, the following linear relationship was
294 obtained in mitochondria, chloroplasts, and chromosomes (Fig.6a), and viruses (Fig.
295 6b): $(X - Y)/(X + Y) = a(X - Y) + b$, where X and Y are nucleotide contents, and (a)
296 and (b) are constants. As (b) was almost null and (a) was ~ 2.0 , $(X - Y)/(X + Y) \approx 2.0(X$
297 $- Y)$. In these genome analyses, which are independent of Chargaff's parity rules (S8
298 Fig., left panels), the values of (a) for (G, C), (G, A), (G, T), (C, T), (C, A) and (A, T)
299 were 2.5858, 1.85558, 1.9908, 1.9771, 1.9968 and 1.5689, respectively. Based on these
300 results, (G + C), (G + A), (G + T), (C + A), (C + T) and (A + T) were 0.39, 0.54, 0.50,
301 0.51, 0.50 and 0.64, respectively. In virus genome analyses (S8 Fig., right panels), the
302 constant values for (a) were 1.9–2.1, and the values for $(X + Y)$ were 0.47–0.53. In
303 contrast, in the normalization of nucleotide contents ($G + C + A + T = 1$), as ($G = C$)
304 and ($A = T$) based on Chargaff's parity rules, ($2G + 2A = 1$) is obtained. This equation is
305 altered to ($G + A = 0.5$). This value is consistent with the value obtained above from
306 genome analyses. Similarly, ($G + T = 0.5$), ($C + A = 0.5$) and ($C + T = 0.5$), although (G

307 + C) and (A + T) cannot be determined. Therefore, the four nucleotide contents are
 308 expressed by the following regression lines, plotted against G content: $A = 0.5 - G$, $T =$
 309 $0.5 - G$, $C = G$ and $(G = G)$. Lines G and C overlap, as do lines A and T, and the former
 310 line is symmetrical to the latter against line ($y = 0.25$). The intercepts of lines G and C
 311 are close to the origin, while those of lines A and T are close to 0.5 at the vertical and
 312 horizontal axes. All organisms from bacteria to *H. sapiens* are located on the diagonal
 313 lines of a 0.5 square, termed the “Diagonal Genome Universe”, using the normalized
 314 values that obey Chargaff’s parity rules [27]. These relationships lead to $(G \text{ or } C) + (A$
 315 $\text{ or } T) = 0.5$. The present results indicate that a linear regression line equation, $(X -$
 316 $Y)/(X + Y) = a(X - Y) + b$, universally represents all normalised values, including the
 317 values deviating from Chargaff’s parity rules. This newly discovered equation clearly
 318 reflects not only Chargaff’s parity rules, based on hydrogen bonding between two
 319 nucleotides, but also natural rule.

320 A linear regression line was not obtained when using randomly chosen value
 321 (Fig. 6c). Furthermore, plotting $(X - Y)/(X + Y)$ against (X/Y) , the following
 322 logarithmic function was obtained for all tested genomes as well as when using
 323 randomly chosen values (Fig. 6 a'-c'): $(X - Y)/(X + Y) = a \ln(X/Y) + b$. As (b) was
 324 almost null and (a) was ~ 0.5 , $(X - Y)/(X + Y) \approx 0.5 \ln(X/Y)$. The ratio between two

values, (X/Y) , can be expressed by a logarithmic function, $\sim 0.5 \ln (X/Y) \approx (X - Y)/(X + Y)$. 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, although non-animal mitochondria, chloroplasts, and chromosomes are controlled under Chargaff's parity rule [17].

333

Fig. 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.

338

339 Conclusions

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

343 biological evolution, can be expressed definitively by a linear regression line: $(X -$
344 $Y)/(X + Y) = a (X - Y) + b$, where $(a \approx 2.0)$ and $(b \approx 0)$ are constants. $(X + Y) =$
345 approximately 0.5. The nucleotide difference, $(X - Y)$, is generally expressed by
346 a linear regression line that crosses “0” at $(X = Y)$, representing no evolution.
347 In addition, the relationship between the ratio (X/Y) and $(X - Y)/(X + Y)$ can be
348 expressed definitively by a logarithmic function, $(X - Y)/(X + Y) = a \ln (X/Y) + b$,
349 where $(a \approx 0.5)$ and $(b \approx 0)$ are constants. The relationship between the skew “ $(X -$
350 $Y)/(X + Y)$ ” and the ratio (Y/X) is universally expressed by a logarithmic line
351 that crosses “1” at $(X = Y)$, representing no change. As the left sides, $(X -$
352 $Y)/(X + Y)$, are equal in both equations, $2.0 (X - Y) \approx 0.5 \ln(X/Y)$. Finally, $(X -$
353 $Y) \approx 0.25 \ln(X/Y)$ using the normalized values from viruses to *H. sapiens* (Fig.
354 7).

355

356 **Fig. 7. Universal rule governing genome evolution from viruses to *H. sapiens*.**

357 Upper pane: cellular organelles, middle panel: viruses, and lower panel: cellular
358 organelles plus viruses.

359

360 **References**

361

362 1. Tsuji G, Fujii S, Sunami T, Yomo T. Sustainable proliferation of liposomes

363 compatible with inner RNA replication. *Proc Natl Ac Sci U S A*. 2016; 113:

364 590-595.

365 2. Schopf JW, Barghoorn ES. Alga-like fossils from the early precambrian of South

366 Africa. *Science*. 1967; 3774: 508-512.

367 3. Nagy B, Zumberge JE, Nagy LA. Abiotic, Graphitic Microstructures in Micaceous

368 Metaquartzite about 3760 Million Years Old from Southwestern Greenland:

369 Implications for Early Precambrian Microfossils. *Proc Natl Acad Sci U S A*. 1975;

370 72: 1206-1209.

371 4. Noffke N, Christian D, Wacey D, Hazen RM. Microbially induced sedimentary

372 structures recording an ancient ecosystem in the ca. 3.48 billion-year-old Dresser

373 Formation, Pilbara, Western Australia. *Astrobiology*. 2013; 13: 1103-1124.

374 5. Schopf JW. Fossil evidence of Archaean life. *Philos Trans R. Soc Lond B Biol*

375 *Sci*. 2006; 361: 869-85.

376 6. Gray MW. The endosymbiont hypothesis revisited. *Int Rev Cytol*. 1992; 141:

377 233-357.

- 378 7. Ku C, Nelson-Sathi S, Roettger M, Sousa FL, Lockhart PJ, Bryant D, et al.
379 Endosymbiotic origin and differential loss of eukaryotic genes.
380 Nature. 2015; 524: 427-32.
- 381 8. Andersson SG, Zomorodipour A, Andersson JO, Sicheritz-Pontén T, Alsmark
382 UC, Podowski RM, et al. The genome sequence of *Rickettsia prowazekii* and the
383 origin of mitochondria. Nature. 1998; 396: 133-140.
- 384 9. Thrash JC, Boyd A, Huggett MJ, Grote J, Carini P, Yoder RJ, et al. Phylogenomic
385 evidence for a common ancestor of mitochondria and the SAR11 clade. Sci
386 Rep. 2011;1:13. doi: 10.1038/srep00013.
- 387 10. Sorimachi K, Okayasu T. Classification of non-animals and invertebrates based on
388 amino acid composition of complete mitochondrial genomes. Intl J Biol. 2014; 6:
389 1-16.
- 390 11. Sorimachi K. Origine of life in the ocean: direct derivation of mitochondria from
391 primitive organisms based on complete genomes. Cur Chem Biol. 2015; 9: 23-35.
- 392 12. Sorimachi K. Genome data provides simple evidence for a single origin of life. Natl
393 Sci. 2010; 2: 519-25.

- 394 13. Sorimachi K, Okayasu T, Ohhira S, Fukasawa I, Masawa N. Evidence for the
395 independent divergence of vertebrate and high C/G ratio invertebrate mitochondria
396 from the same origin. Natl Sci. 2012; 4: 479-483.
- 397 14. Sorimachi K. Evolutionary changes reflected by the cellular amino acid composition.
398 Amino Acids. 1999; 17: 207-226.
- 399 15. Sorimachi K, Itoh T, Kawarabaysi Y, Okayasu T, Akimoto K, Niwa A.
400 Conservation of the basic pattern of cellular amino acid composition during
401 biological evolution and the putative amino acid composition of primitive life forms.
402 Amino Acids. 2001; 21: 393-399.
- 403 16. Chargaff E. Chemical specificity of nucleic acids and mechanism of their enzymatic
404 degradation. Experientia. 1950; 6: 201-209.
- 405 17. Rundner R, Karkas JD, Chargaff E. Separation of *B. subtilis* DNA into
406 complementary strands. 3. Direct analysis. Proc Natl Acad Sci USA. 1968; 60:
407 921-922.
- 408 18. Sorimachi K. A proposed solution to the historic puzzle of Chargaff's second parity
409 rule. Open Genom J. 2009; 2: 12-14.
- 410 19. Sorimachi K, Okayasu T, Ohhira S. Normalization of complete genome
411 characteristic: Application to evolution from primitive organisms to *Homo sapiens*.
412 Cur Genom. 2015; 16: 99-106.
- 413 20. Sorimachi K. Indices of whole genome characteristics based on amino acid or

- 414 nucleotide content ratios: evolution from primitive organisms to *Homo sapiens*. J
415 Chem Engineer Chem Res. 2014; 1: 302-312.
- 416 21. Mitchell D, Bridge R. A test of Chargaff's second rule. Biochem Biophys Res
417 Commun. 2006; 340: 90-94.
- 418 22. Sorimachi K, Okayasu T. Codon evolution is governed by linear formulas. Amino
419 Acids. 2008; 34: 345-350.
- 420 23. Sorimachi K, Okayasu T. Universal rules governing genome evolution expressed by
421 linear formulas. Open Cur Genom. 2008; 1: 33-43.
- 422 24. Nikolaou C, Almirantis Y. Deviations from Chargaff's second parity rule in
423 organellar DNA Insights into the evolution of organellar genomes. Gene. 2006; 381:
424 34-41.
- 425 25. Bell SJ, Forsdyke DR. Deviations from Chargaff's second parity rule correlate with
426 direction of transcription. J Theor Biol. 1999; 197] 63-76.
- 427 26. Sorimachi K, Okayasu T. Gene assembly consisting of small units with similar
428 amino acid composition in the *Saccharomyces cerevisia*. Mycoscience. **2003**; 44:
429 415-417.
- 430 27. Sorimachi K. Evolution based on genome structure: the "diagonal genome universe".
431 Natl Sci. 2010; 2: 1104-1112.
- 432 28. Lobry J R. Asymmetric substitution patterns in the two DNA strands of bacteria.
433 Mol. Biol. Evolution. 1996; 13: 660-665.

- 434 29. Tillier E R, Collins R A. The contributions of replication orientation, gene direction,
435 and signal sequences to base-composition asymmetries in bacterial genomes. J Mol
436 Evo. 2000; 50: 249-257.
- 437 30. Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, et al.
438 Sequence and organization of the human mitochondrial genome. Nature. 1981; 290:
439 457-65.
- 440 31. Fonceca MM, Harris DJ, Posada D. The inversion of the control region in three
441 mitogenomes provides further evidence for an asymmetric model of vertebrate
442 mtDNA replication. PloS One. 2014; 9, e106654.
- 443 32. Seligmann H. Coding constraints modulate chemically spontaneous mutational
444 replication gradients in mitochondrial genomes. Cur Geno. **2012**; 13: 37-54.
- 445 33. King N, Westbrook MJ, Young SL, Kuo A, Abedin M, Chapman J, et al. The
446 genome of the choanoflagellates *Monosiga brevicollis* and the origin of metazoans.
447 Nature. 2008; 451: 783-788.
- 448 34. Groseth A, Stroher U, Theriault S, Feldmann H. Molecular characterization of an
449 isolate from the 1989/90 epizootic of Ebola virus Reston among macaques imported
450 into the United States. Virus Res. **2002**; 87: 155-163.
- 451 35. Sanchez A, Rollin PE. Complete genome sequence of an Ebola virus (Sudan
452 species) responsible for a 2000 outbreak of human disease in Uganda. Virus Res.
453 2005; 113: 16-25.

- 454 36. Volchkov VE, Volchkova VA, Chepurnov AA, Blinov VM, Dolnik O, Netesov SV,
455 et. Characterization of the L gene and 5' trailer region of Ebola virus. J Gen Virol.
456 1999; 80 (Pt 2): 355-362.
- 457 37. Towner JS, Sealy TK, Khristova ML, Albariño CG, Conlan S, Reeder SA, et al.
458 Newly Discovered Ebola Virus Associated with Hemorrhagic Fever Outbreak in
459 Uganda. *PLoS Pathog.* 2008; 4(11): E1000212 doi: 10.1371/journal.ppat.1000212.
- 460 38. Jenkins PD, Kilpatorick CW, Robinson MF, Timmins RJ. Morphological and
461 molecular investigations of a new family, genus and species of rodent (Mammalia:
462 Rodentia: Hystricognatha) from Lao PDR. *Syst Biodiv.* 2004;2: 419-454.
- 463 39. Dawson MR, Marivaux L, Li CK, Beard KC, Métais G. Laonastes and the "Lazarus
464 effect" in recent mammals. *Science.* 2006; 311: 1456-1458.
- 465 40. Gaudin TJ, Emry RJ, Wible JR.: The phylogeny of living and extinct pangolins
466 (Mammalia, Pholidota) and associated taxa: a morphology based analysis. *J Mam*
467 *Evol.* 2009; 16: 235-305.
- 468 41. Gatesy J, Hayashi C, Cronin MA, Arctander P. Evidence from milk casein genes
469 that cetaceans are close relatives of hippopotamid artiodactyls. *Mol Biol Evol.* 1996;
470 13: 954-963.
- 471 42. Ursing BM, Arnason U. Analyses of mitochondrial genomes strongly support a
472 hippopotamus-whale clade. *Proc Biol Sci.* 1998; 265: 2251-2255.

- 473 43. Sorimachi K, Okayasu T. Phylogenetic tree construction based on amino acid
474 composition and nucleotide content of complete vertebrate mitochondrial genomes.
475 IOSR J. Pharm. 20013; 3: 51-60.
- 476 44. Session AM. , **Uno Y, Kwon T, Chapman JA, Toyoda A, Takahashi S**,et al.
477 Genome evolution in the allotetraploid frog *Xenopus laevis*.
478 Nature. 2016; 538: 336-343.
- 479 45. Hellsten, U. Simakov O, Chapman J, Fahey B, Gauthier ME, Mitros T, et al. The
480 genome of the Western clawed frog *Xenopus tropicalis*.
481 Science. 2010; 328: 633-636.
- 482 46. Pollet N, Mazabraud A. Insights from *Xenopus* genomes. Genome Dyn. 2006; 2:
483 138-153.
- 484 47. Tymowska J. Karyotype analysis of *Xenopus tropicalis* Gray, Pipidae. Cytogenet.
485 Cell Genet. 1973; 12: 297-304.
- 486 48. Sakaguchi DS, Murphey RK, Hunt RK, Tompkins R. The development of retinal
487 ganglion cells in a tetraploid strain of *Xenopus laevis*: a morphological study
488 utilizing intracellular dye injection. J Comp Neurol. 1984; 224: 231-251.
- 489 49. Ohno S. Evolution by Gene Duplication (Springer, 1970)
- 490 50. Janvier P. Micro RNAs revive old views about jawless vertebrate divergence and
491 evolution. Proc Natl Acad Sci USA. 2010; 107: 19137-1918.

492 51. van Rheede T, Bastiaans T, Boone DN, Hedges SB, de Jong WW, Madsen O. The
493 platypus is in its place: nuclear genes and indels confirm the sister group relation of
494 monotremes and Therians. Mol Biol Evol. 2006; 23: 587-597.

495

496 **Supporting information**

497 **S1 Fig. Nucleotide content differences in the three fragments of the mitochondrial**
498 **genomes of invertebrates.** Left to right: (G – C), (G – T), (G – A), (C – T), (C – A), and
499 (T – A).

500 **S2 Fig. Nucleotide content differences in the three fragments of the mitochondrial**
501 **genomes of vertebrates.** Left to right: (G – C), (G – T), (G – A), (C – T), (C – A), and
502 (T – A).

503 **S3 Fig. Nucleotide content differences in the three fragments of the mitochondrial**
504 **genomes of non-animal mitochondria and chloroplasts.** Left to right: (G – C), (G –
505 T), (G – A), (C – T), (C – A), and (T – A).

506 **S4 Fig. Nucleotide content differences in the three fragments of bacterial genomes.**
507 Left to right: (G – C), (G – T), (G – A), (C – T), (C – A), and (T – A).

508 **S5 Fig. Nucleotide content differences in the three fragments of the chromosome of**
509 **microsporidian protozoan species *Encephalitozoon cuniculi*.** Left to right: (G – C),

510 (G – T), (G – A), (C – T), (C – A), and (T – A).

511 **S6 Fig. Nucleotide content differences in the three fragments of viral genomes.** Left

512 to right: (G – C), (G – T), (G – A), (C – T), (C – A), and (T – A).

513 **S7 Fig. Nucleotide differences.** (a) Mitochondria of aquatic vertebrates (blue arrow)

514 and terrestrial vertebrates (red arrow). (b) Mitochondria of high C/G invertebrates

515 (black arrow) and low C/G invertebrates (red arrow). (c) Non-animal mitochondria of

516 fungi (blue arrow) and plants (green arrow). (d) Chloroplasts. (e) Chromosomes of

517 prokaryotes (blue arrow), archaea (green arrow), and eukaryotes (red arrow).

518 **S8 Fig. Universal rules.** Left side: relationship between (X – Y) and (X – Y)/(X + Y),

519 and right side: relationship between (X/Y) and (X – Y)/(X + Y). (a): organelles and (b):

520 viruses.

521 **S1 Table. List of viruses used.**

522 **S1 Text. Application of newly developed analytical method to various organisms'**

523 **mitochondrial genomes.**

524

a **b** **c** **d** **e** **f**

a*Ureaplasma urealyticum* **d***Mycosiga brevicollis* Mt **b** **e** **c** **f**

Mnosiga brevicollis

Mnosiga brevicollis

Polysphondyliumpallidum

Polysphondyliumpallidum

Dictyosteliumdiscaideum

Dictyosteliumdiscaideum

Reclinomonas americana

Reclinomonas americana

Homo sapiens

Daubentonia madagascariensis

Pan troglodytes

Nycticebus coucang

Gorilla gorilla

Tupaia belangeri

Macaca mulatta

Lemur catta

Ebola virus Reston

Ebola virus Bundibugyo

Ebola virus Sudan

Ebola virus Tai Forest

Ebola virus Zaire

Mitochondria, Chloroplasts, Chromosomes

Viruses

Random numbers