Ancient DNA sequence revealed by error-correcting codes

Marcelo M. Brandão, Larissa Spoladore, Luzinete C. B. Faria, Andréa S. L. Rocha, Marcio C. Silva-Filho & Reginaldo Palazzo Jr.

A previously described DNA sequence generator algorithm (DNA-SGA) using error-correcting codes has been employed as a computational tool to address the evolutionary pathway of the genetic code. The code-generated sequence alignment demonstrated that a residue mutation revealed by the code can be found in the same position in sequences of distantly related taxa. Furthermore, the code-generated sequences do not promote amino acid changes in the deviant genomes through codon reassignment. A Bayesian evolutionary analysis of both code-generated and homologous sequences of the *Arabidopsis thaliana* malate dehydrogenase gene indicates an approximately 1 MYA divergence time from the MDH code-generated sequence node to its paralogous sequences. The DNA-SGA helps to determine the plesiomorphic state of DNA sequences because a single nucleotide alteration often occurs in distantly related taxa and can be found in the alternative codon patterns of noncanonical genetic codes. As a consequence, the algorithm may reveal an earlier stage of the evolution of the standard code.

Biological and digital communication systems have similarities with respect to the corresponding procedures used to convey the biological and digital information from one point to another, as well as in the data storage of digital media in a redundant array of independent disks (RAID) and the storage of genetic information in chromosomes. These similarities enable the use of algorithms in the modeling and analyses of biological systems and data. For instance, in eukaryotic cells, the information contained in the DNA is transmitted through RNA to produce the proteins needed at a precise moment and in specific compartments in the cell. Many enzymes and complex molecules coordinate their transport and are often assisted by protein intermediates in the cytosol and organellar membranes, thus identifying the correct location of a protein. In the same way, the transmission of flawless data through noisy channels in digital communication systems can be reliably achieved if, in addition to using an error-correcting code (ECC), extensive signal processing techniques are also employed.

For quite some time there have been attempts to confirm the existence of an error-control mechanism in biological sequences similar to the ECC employed in digital sequences, and although relevant, such studies have yet to provide a definitive answer. Recently our group developed an algorithm, known as DNA Sequence Generator Algorithm, which verifies whether a given DNA sequence can be identified as a codeword of a G-linear code. This goal was achieved when many distinct DNA sequences were identified as code words of G-linear codes (consisting of specific mappings and the underlying BCH codes) an important subclass of cyclic codes.

BCH codes were first proposed by Hocquenghem and independently rediscovered by Bose and Chaudhuri; therefore, the acronym is made up of the initials of Bose, Chaudhuri, and Hocquenghem. When an underlying BCH code over Galois ring extension and/or Galois field extension identifies a given DNA sequence, two things may occur: 1) the given DNA sequence is a codeword of a G-linear code.
code; or 2) it is a sequence belonging to the set of neighboring sequences differing by at least one nucleotide from the corresponding codeword of a G-linear code. This set of neighboring sequences is referred to as the “cloud” of a codeword.

When the DNA sequence generation algorithm identifies a DNA sequence belonging to the cloud of a codeword, it differs in a single nucleotide from the original sequence. Similar to biological DNA, this generated codeword may represent a silent mutation causing no effect on the translated amino acid or it may cause a residue change affecting for instance the protein structure and activity and consequently impairing its interactions with other proteins. Furthermore, the single nucleotide alteration can be restored, or equivalently, the codeword can be reverse engineered, returning it to its original sequence by applying one of the following algorithms: the Berlekamp-Massey decoding algorithm for codes over Galois field extensions or the Modified Berlekamp-Massey decoding algorithm for codes over Galois ring extensions, together with the corresponding labeling associated with each analyzed sequence.

Recently, Ivanova and colleagues used a metagenomics approach to survey the prevalence of stop codon reassignment in naturally occurring microbial populations and proposed that the canonical genetic code may contain some deviations. Similarly, studies of the evolution of the genetic code have developed a hypothesis that differs from a frozen universal code and even the universality of the code. It has been observed that each deviant genetic code contains codons that are associated with different amino acids and also with the canonical genetic code. Consequently, one may infer that such a process may have evolved from a standard code. Such deviant genetic codes can be found in nuclear and mitochondrial genomes, in which mechanisms of codon reassignment have led to the differential reading of certain codons. The evolution of the genetic code plays an important role in understanding the differences between the response of the DNA sequence identification process and the given DNA sequence because these differences can be related to either the canonical genetic code or to the several deviant genetic codes. In another example, Inomata and colleagues using multiple sequence alignment and test of neutrality, have demonstrated that a single replacement of guanine with adenine (position 2 of the gene) in Drosophila melanogaster, resulting on threonine at the 218 - amino acid position, was the ancestral form of the Gr5a gene in D. melanogaster and this single amino acid polymorphism (ALA218THR) represents a key impact on the trehalose sensitiveness.

The proposal of mathematical models describing such biological systems provides the needed tools for the development of systematic approaches for studies of mutations and polymorphisms and has applications in genetic engineering.

Thus, if an ECC can identify differences in a DNA sequence with a one-nucleotide resolution, the questions that should be addressed are as follows: if there is an error-correcting code underlying the DNA sequences, what are the biological implications regarding the single nucleotide (SNP) difference? And, is there a biological reasoning for such a difference? In the present study, we used the ECC approach proposed in references to evaluate whether the nucleotide difference between the original DNA sequence and the sequence identified as the codeword of the ECC is biologically significant in terms of evolution of this identified polymorphism.

Results and Discussion

In this study the DNA sequence generation algorithm was applied as a computational tool to provide strong evidence of the evolution of the genetic code, in special on nucleotide and amino acid site specific polymorphism, by showing the existence of a mathematical structure underlying the actual DNA sequences and by investigating the real biological meaning of the difference in the specific position pointed out by the code-generated sequences.

The code-generated sequences that had a single nucleotide alteration, causing a residue change in the translated protein, were used in a Blastx analysis to verify if the alteration suggested by the ECC could be found in other sequences.

Analyses were run for the code generated sequences of the Saccharomyces YMR193 gene (GI 45269853), the Triticum aestivum wPR4 gene (GI 78096542), the Nicotiana tabacum antifungal CBP 20 gene (GI 632733), the Citrus sinensis chlorophyllase gene (GI 7328566), the Arabidopsis thaliana heme-like protein PR4 gene (GI 186509758), the Saccharomyces cerevisiae OXA gene (GI 832917) and the Homo sapiens F1F0 ATP-synthase gene (GI 12587). These sequences are shown in Table S01.

As a result of this search approach, a number of different genes were found to contain the same nucleotide at the same altered position suggested by the error-correcting code, see Tables 1 and S02.

In some of the results, the suggested polymorphism could be found in DNA sequences of taxa that were closely related to the query sequence. For example, for the code-generated sequence of the YMR 193 gene from Saccharomyces cerevisiae (Tables S02a and S02b), a mitochondrial protein involved in the large ribosomal subunit, the same residue was also found in other Ascomycota sequences. The results for the code-generated antifungal CBP 20 gene from Nicotiana tabacum were similar to those from other eudicots. The results for the chlorophyllase gene in Citrus sinensis were also found in sequences in Populus spp. And the code-generated sequence for the OXA gene, which is involved in cytochrome oxidase biogenesis in Saccharomyces cerevisiae, showed the same residue in the altered position in other ascomycete sequences. However, different cases were found as well, such as the F1F0 ATP-synthase gene from Homo sapiens, in which the code-generated polymorphism of His to Gln could only be found at the same position in certain fungi sequences, a very distantly related taxa to H. sapiens, which may
by the Klein-linear code ((1023, 1013, 3) BCH code over \(\mathbb{Z}_4\) with the generator polynomial \(g(x)\). The analyses showed that the malate dehydrogenase sequences form a monophyletic group rooted in the sequence generated by the ECC (Fig. 1). This sequence was generated for the code-generated and homologous sequences of the Arabidopsis thaliana malate dehydrogenase gene. The analyses showed that the Arabidopsis thaliana malate dehydrogenase gene sequence in Triticum aestivum (GI 217937) was applied to the DNA Sequence Table 1. *Polynomial-based DNA sequences generated by BCH codes over Galois ring and field extensions. Abbreviations: TS targeting sequence; PM protein motifs; L labelings A, B, C and D; R ring; F field; *signal or transit peptide without experimental evidence. Sc Saccharomyces cerevisiae; Ts Nicotiana tabacum; Cs Citrus sinensis; At Arabidopsis thaliana; Hs Homo sapiens; Ib Ipomoea batatas; Pd Polistes dominulus; Mm Mesobuthus martensi.*

instead provide evidence that this algorithm may be describing ancient site specific sequences in which evolution acted to influence the current appearance of the gene. This was also observed in the wPR4 gene, which is involved in vacuolar defense in Triticum aestivum, in which the residue alterations in the evolution acted to influence the current appearance of the gene. This was also observed in the wPR4 gene sequences. These observations suggest that the sequence generated by the code might be more closely related to the ancestor of Arabidopsis thaliana malate dehydrogenase rather than to other paralogous genes, evidencing that the ECC generated sequence has a SNP that may be indicating the ancient state of this sequence. The application of an ECC does not aim to reconstruct full ancestral sequences from a given phylogenetic tree and aligned gene sequences of some current species; here we describe an ancestral site specific reconstruction based solely on DNA primary structure recovered from coding and decoding gene sequences.

Among the analyzed sequences, we identified several single nucleotide polymorphisms that were pointed out by the ECC as leading to a codon alteration (and also an amino acid alteration in the translated sequence), but in the deviant genetic codes, these altered codons correspond to the same amino acids that were found in the original sequence.15,26,27.

In noncanonical genetic codes, alterations in the components of the translation mechanism confer different meanings to specific codons. For example, TGA is read as Trp 17,27–38, AGA as Ser 33–37, ATA as Ser 17,27, and TGA as Cys17,23.

When the F1 ATPase gene from Ipomoea batatas (GI 217937) was applied to the DNA Sequence Generator Algorithm, the output sequence presented an alteration in the sense codon TGG (encoding Trp) to become the stop codon TGA (Table 4a). A similar example is observed in the BRCA1 gene sequence in H. sapiens (GI 25140446), which is altered by the code from Cys to a stop codon (Table 4b). Interestingly, in the mitochondrial genetic code of most organisms, aside from green plants, the codon TGA is associated with tryptophan, and studies have shown that in the primary structure of
mitochondrial and nuclear genomes, the TGA codon does not signal for the release of the transcription factors but instead codes for Trp. The code generated sequence for the Allergen Pol d5 gene of *Polistes dominulus* (GI 51093376) showed an alteration from AGT (Ser) to AGA (Arg), the same happening with the code generated sequence for the anti-epilepsy peptide precursor of *Mesobuthus martensii* (GI 16740522) showed an alteration from ATG (Met) to ATA (Ile) (Table 4b,c). In noncanonical genetic codes, AGA codes for Ser and ATA for Met; therefore, these alterations could modify the folding and activity of the subsequent protein due to a change in the charge and hydrophathy of the residues.

Often times, the same codon reassignment may independently occur multiple times in different taxa. The mechanisms leading to codon reassignment have yet to be fully elucidated and may be due to factors such as codon disappearance, an ambiguous intermediate, or unassigned codons. Deviant genetic codes are an example of how populations cross over maladaptive valleys from one adaptive peak to another, in respect to error minimization, via adaptive bridges. Therefore, the algorithm may underlie any of the stages of information transmission, representing an earlier stage of the evolution of the universal/canonical code.

The characters, character states, and the evolution of ancient genes or proteins can hardly be directly studied, because such molecule are rarely preserved over the evolutionary time or from any ancestral, living or preserved, has not been gathered from the nature. Pauling and Zuckerkandl once proposed that ancestral molecules could one day be “resurrected” by digging out from the evolution their ancient form. Since then, different methods of ancestral sequence reconstruction (ASR) have emerged based on parsimony, Bayesian inference or maximum likelihood. Independently of the methodology used all these approaches rely on multiple sequence alignment with the aim of elucidating the complete and distant sequences (Supplementary material 1 presents a maximum likelihood analysis for the *Arabidopsis thaliana* Malate Dehydrogenase). Here, we hypothesize that the G-linear code may identify the original molecular primary structure of the sequence using only the intrinsic nucleotide composition. The DNA sequence generation algorithm can describe the plesiomorphic state of certain DNA character state sequences, as the suggested single nucleotide alteration often occurs in distant taxa and is maintained by alternative codon patterns in noncanonical genetic codes.

In summary, the G-linear code, commonly associated with reliable digital transmission, even with all the constraints inherent to the construction of the ECC, unwraps the molecular component of every living cell when it is applied to the primary structure of DNA, thus revealing ancient information that may have been silenced by assorted evolutionary pressures that have shaped the present forms of life. This code generates point mutations that can be found in actual (real) sequences, and the DNA sequence generation algorithm can be used in computer simulations for the analysis of polymorphisms and mutations.
Methods

Identification of the DNA sequences. Although several DNA-encoding sequences (organelle-targeting sequences, introns, protein motifs, and full proteins) were identified by the corresponding G-linear codes over finite Galois rings and fields, as shown in Table 1, the majority of these DNA sequences were identified by the G-linear codes over rings. One possible explanation is that the latter algebraic structure may be more flexible than the algebraic structure of fields. As a consequence, the sequences identified by the corresponding G-linear codes over fields exhibit less adaptability than those offered by G-linear codes over rings. This observation suggests that it is possible to classify the proteins according to their nature.

Table 2. *A. thaliana* - Mitochondrial - Malate dehydrogenase 1 - GI number 30695458.
stability in the mutation index, allowing a new approach for the classification of DNA sequences from a mathematical point of view.

All of the DNA sequences analyzed by the DNA sequence generation algorithm were identified as belonging to the “cloud” of the corresponding code words of the ECC. In other words, the actual DNA sequences differ from the corresponding code words of the ECC by a single nucleotide. The code-generated sequences in which the single nucleotide alteration led to an amino acid change in the translated protein were further analyzed.

These code-generated sequences were used as queries in a Blastx search, with the results filtered for green plants, fungi, bacteria, Archaea, algae and monocots from the NCBI non-redundant protein sequence database. The Blastx results were then aligned with Muscle50,51 (CLC Bio Genomics workbench plugin) and the position of the altered amino acid was compared with these results.

Several codons with the same meaning have been reassigned in independent lineages, which could mean that there is an underlying predisposition towards certain reassignments43. As an example of how

| A. thaliana | Mitochondrial - Malate dehydrogenase 1 | GI number 30695458 |
|-------------|----------------------------------------|-------------------|
| Abbreviations: | | |
| Oaa = original amino acid, Ont = original nucleotide, Olb = original labeling, Glb = generated labeling, Gnt = generated nucleotide, Gaa = generated amino acid. Red: shows where the error occurred in the targeting sequences. p(x) = primitive polynomial; p(x)' = reciprocal polynomial of p(x). g(x) = generator polynomial; g(x)' = reciprocal generator polynomial of g(x). | | |

Table 3. A. thaliana - Mitochondrial - Malate dehydrogenase 1 – GI number 30695458. Abbreviations: Oaa = original amino acid, Ont = original nucleotide, Olb = original labeling; Glb = generated labeling; Gnt: generated nucleotide; Gaa: generated amino acid. Red: shows where the error occurred in the targeting sequences. p(x) = primitive polynomial; p(x)' = reciprocal polynomial of p(x). g(x) = generator polynomial; g(x)' = reciprocal generator polynomial of g(x). | | |
a) *L. batatas* – Mitochondrial – F1ATPase delta subunit – GI number 2179377

**G-linear code** \(((63,57,3)) \) Primitve BCH code over \(GF(64)\), labeling \(D=(A=0,C=1,G=a,T=b)\)

| Cnt: | ATG | TTT | AGG | CAC | TCT | CGA | CTC | CTA | GCT | GCC | GCC | ACC | ACA | ATG | GGG | TGG | CGT | CGC | CCC | TTC |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Glb: | 0bb | bbb | aab | 1bb | 0aa | 1ab | bab | 0ba | bab | 1bb | 0aa | 1ab | bab | 0ba | bab | 1bb | 0aa | 1ab | bab | 0ba | bab |
| Gnt: | ATG | TTT | AGG | CAC | TCT | CGA | CTC | CTA | GCT | GCC | GCC | ACC | ACA | ATG | GGG | TGG | CGT | CGC | CCC | TTC |
| Gaa: | M | F | R | S | R | L | A | R | A | T | T | M | G | sto | R | R | P | F | F |

**Evolutionary proposal and estimation of the divergence time based on Bayes approach**

Estimates of divergence time among malate dehydrogenase \((\text{MDH})\) sequences.

- The divergence time between fungi and green plants\(^{52}\), mosses and vascular plants\(^{52,53}\) and eudicot rosids and asterids\(^{54}\) was used to estimate a divergence time for the *Arabidopsis thaliana* group. Species-level phylogenies were generated using a Bayesian uncorrelated lognormal relaxed clock model in Beast version 1.4.8\(^{55}\). The dataset followed the GTR + \(\Gamma\) model of substitution implemented in Beast, and two Monte Carlo Markov chains were run for 90,000,000 generations, using the Yule speciation model, using a 10% burn-in with sampling trees generated every 10,000 generations.

b) *P. dominulus* – Endoplasmic reticulum – Allergen Pol d 5 – GI number 5109376

**G-linear code** \(((63,57,3)) \) Primitve BCH code over \(GF(64)\), labeling \(D=(A=0,C=1,G=a,T=b)\)

| Cnt: | ATG | AAA | AGG | TTA | TTA | TTA | CTA | GAT | TCT | GCT | GCC | GCC | ACC | ACA | ATG | GGG | TGG | CGT | CGC | CCC | TTC |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Glb: | 0bb | 0aa | bab | 1bb | 0aa | 1ab | bab | 0ba | bab | 1bb | 0aa | 1ab | bab | 0ba | bab | 1bb | 0aa | 1ab | bab | 0ba | bab |
| Gnt: | ATG | AAA | AGG | TTA | TTA | TTA | CTA | GAT | TCT | GCT | GCC | GCC | ACC | ACA | ATG | GGG | TGG | CGT | CGC | CCC | TTC |
| Gaa: | M | K | I | R | C | I | C | L | V | I | V | L | T | I | H | L | S | Q | A |

Evolutionary proposal and estimation of the divergence time based on Bayes approach

Estimates of divergence time among malate dehydrogenase \((\text{MDH})\) sequences.

- The divergence time between fungi and green plants\(^{52}\), mosses and vascular plants\(^{52,53}\) and eudicot rosids and asterids\(^{54}\) was used to estimate a divergence time for the *Arabidopsis thaliana* group. Species-level phylogenies were generated using a Bayesian uncorrelated lognormal relaxed clock model in Beast version 1.4.8\(^{55}\). The dataset followed the GTR + \(\Gamma\) model of substitution implemented in Beast, and two Monte Carlo Markov chains were run for 90,000,000 generations, using the Yule speciation model, using a 10% burn-in with sampling trees generated every 10,000 generations.

c) *M. martensi* – Endoplasmic reticulum – anti-epilepsy peptide precursor – GI number 16740522

**Z-linear code** \(((63,57,3)) \) Primitve BCH code over \(GF(4,6)\), labeling \(A=(A=0,C=1,G=3,T=2)\)

| Cnt: | ATG | AAA | CTA | TTT | TTT | TTA | TTA | CTA | GAT | GCT | TCT | TCA | ATG | GTA | GAT | GGC | TTA | GCT | ATT | TGG | CGC | CCC | TTC |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Glb: | 0bb | 0aa | bab | 1bb | 0aa | 1ab | bab | 0ba | bab | 1bb | 0aa | 1ab | bab | 0ba | bab | 1bb | 0aa | 1ab | bab | 0ba | bab |
| Gnt: | ATG | AAA | CTA | TTT | TTT | TTA | TTA | CTA | GAT | GCT | TCT | TCA | ATG | GTA | GAT | GGC | TTA | GCT | ATT | TGG | CGC | CCC | TTC |
| Gaa: | M | K | L | F | L | L | L | V | I | S | A | S | A | S | A | S | A | S | A | S | A | S | A |

d) *H. sapiens* – truncated breast and ovarian cancer susceptibility protein \((\text{BRCA1})\) gene, exon 12 and partial cds – GI number 25140446

**Z_{m2}-linear code** \(((63,57,3)) \) Primitve BCH code over \(GF(4,6)\), labeling \(B=(A=0,C=1,G=2,T=3)\)

| Cnt: | GAA | GCA | GCA | TCT | GGC | TGG | TGT | GAG | AGT | GAA | AGA | AGC | GTC | TCT | GAA | GAC | TGC | TCA | GGG | CTA | TCA | GAG |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Glb: | 200 | 210 | 210 | 210 | 210 | 210 | 210 | 210 | 210 | 210 | 210 | 210 | 210 | 210 | 210 | 210 | 210 | 210 | 210 | 210 | 210 | 210 |
| Gnt: | GAA | GCA | GCA | TCT | GGC | TGG | TGT | GAG | AGT | GAA | AGA | AGC | GTC | TCT | GAA | GAC | TGC | TCA | GGG | CTA | TCA | GAG |
| Gaa: | E | A | A | A | S | G | C | E | G | S | T | V | S | E | S | D | C | S | G | L | S |

**Table 4. DNA sequences generated by BCH code.** Abbreviations: Oaa = original amino acid, Ont = original nucleotide, Olb = original labeling; Glb: generated labeling; Gnt: generated nucleotide; Gaa: generated amino acid. Red: shows where the alteration has occurred in the targeting sequences. \(p(x)\) = primitive polynomial ; \(p(x)’\) = reciprocal polynomial of \(p(x)\). \(g(x)\) = generator polynomial ; \(g(x)’\) = reciprocal generator polynomial of \(g(x)\).

The DNA sequence generation algorithm could be determining the ancient codon patterns in the analyzed species, we searched for codons in the code-generated sequences that were related to meaningful biological parts of nuclear or mitochondrial genomes\(^{16,27}\) and had a codon reassignment in other species.

**Evolutionary proposal and estimation of the divergence time based on Bayes approach**

Estimates of divergence time among malate dehydrogenase \((\text{MDH})\) sequences.

- The divergence time between fungi and green plants\(^{52}\), mosses and vascular plants\(^{52,53}\) and eudicot rosids and asterids\(^{54}\) was used to estimate a divergence time for the *Arabidopsis thaliana* group. Species-level phylogenies were generated using a Bayesian uncorrelated lognormal relaxed clock model in Beast version 1.4.8\(^{55}\). The dataset followed the GTR + \(\Gamma\) model of substitution implemented in Beast, and two Monte Carlo Markov chains were run for 90,000,000 generations, using the Yule speciation model, using a 10% burn-in with sampling trees generated every 10,000 generations.
References

1. Patterson, D. A., Gibson, G. & Katz, R. H. A case for redundant arrays of inexpensive disks (RAID). SIGMOD Rec 17, 109–116 (1988).
2. Benedetto, S., Biglieri, E. & Castellani, V. Digital transmission theory. (Prentice-Hall, 1987).
3. MacWilliams, F. J. The theory of error correcting codes / F. J. MacWilliams, N.J.A. Sloane. (North-Holland Pub. Co.; sole distributors for the U.S.A. and Canada, Elsevier/North-Holland, 1977).
4. Faria, L. C. B. et al. Is a genome a codeword of an error-correcting code? PLoS One 7, e36644 (2012).
5. Faria, L. C. B., Rocha, A. S. L. & Palazzo, R. Jr. Transmission of intra-cellular genetic information: A system proposal. Journal of theoretical biology 358, 208–231 (2014).
6. Faria, L. C. B., Rocha, A. S. L., Kleinschmidt, J. H., Palazzo, R. & Silva-Filho, M. C. DNA sequences generated by BCH codes over GF(4). Electronics Letters 46, 203–204 (2010).
7. Rocha, A. S. L., Faria, L. C. B., Kleinschmidt, J. H., Palazzo, R. Jr. & Silva-Filho, M. C. DNA sequences generated by Z4-linear codes. 2010 IEEE International Symposium on Information Theory Proceedings (ISIT), 1320–1324 (2010).
8. Hocquenghem, A. Codes correcteurs de erreurs. Clairefs 2, 127–136 (1959).
9. Bose, R. C. & Ray-Chaudhuri, D. K. On a class of error correcting binary group codes. Information and Control 3, 68–79 (1960).
10. Berlekamp, E. R. Algebraic coding theory. (McGraw-Hill, 1968).
11. Massey, J. L. Shift-Register Synthesis and Bch Decoding. Ieee T Inform Theory 15, 122–127 (1969).
12. Elia, M., Interlando, J. C. & Palazzo, R. Computing the reciprocal of units in Galois rings. Journal of Discrete Mathematical Sciences and Cryptography 3, 41–55 (2000).
13. Interlando, J. C, Palazzo, R. J. & Elia, M. On the decoding of Reed-Salomón and BCH codes over integer residue rings. Ieee T Inform Theory 43, 1013–1021 (1997).
14. Ivanova, N. N. et al. Stop codon reassignments in the wild. Science 344, 909–913 (2014).
15. Crick, F. H. The origin of the genetic code. J Mol Biol 38, 367–379 (1968).
16. Knight, R. D., Freeland, S. J. & Landweber, L. F. Rewiring the keyboard: evolvability of the genetic code. Nat Rev Genet 2, 49–58 (2001).
17. Osawa, S. & Jukes, T. H. Codon reassignment (codon capture) in evolution. J Mol Evol 28, 271–278 (1989).
18. Osawa, S. Z. Evolution of the genetic code. (Oxford University Press, 1995).
19. Yokobori, S., Suzuki, T. & Watanabe, K. Genetic code variations in mitochondria: tRNA as a major determinant of genetic code plasticity. J Mol Evol 53, 314–326 (2001).
20. Jukes, T. H. & Osawa, S. Evolutionary changes in the genetic code. Comp Biochem Physiol B 106, 489–494 (1993).
21. Anderson, S. et al. Sequence and organization of the human mitochondrial genome. Nature 290, 457–465 (1981).
22. Kawahara-Kobayashi, A. et al. Simplification of the genetic code: restricted diversity of genetically encoded amino acids. Nucleic Acids Res 40, 10576–10584 (2012).
23. Lozupone, C. A., Knight, R. D. & Landweber, L. F. The molecular basis of nuclear genetic code change in ciliates. Curr Biol 11, 65–74 (2001).
24. Yoko, T., Okimoto, R. & Wolstenholme, D. R. The mitochondrial genome of the sea anemone Metridium senile (Cnidaria): Introns, a paucity of tRNA genes, and a near-standard genetic code. J Mol Biol 247, 204–211 (1995).
25. Inomata, N. A Single-Amino-Acid Change of the Gustatory Receptor Gene, Gr5a, Has a Major Effect on Trehalose Sensitivity in Drosophila melanogaster. Genetics 170, 969–976 (2014).
26. Sengupta, S., Yang, X. & Higgs, P. G. The mechanisms of codon reassignments in mitochondrial genetic codes. J Mol Evol 64, 662–688 (2007).
27. Swire, J., Judson, O. P. & Burt, A. Mitochondrial genetic codes evolve to match amino acid requirements of proteins. J Mol Biol 60, 126–139 (2005).
28. Hayashishimaru, Y., Ebara, M., Inagaki, Y. & Ohama, T. A deviant mitochondrial genetic code in prymnesiophytes (yellow-algae): UGA codon for tryptophan. Curr Genet 32, 296–299 (1997).
29. Turmel, M. et al. The complete mitochondrial DNA sequences of Nephropselis ovicola and Pedinomonas minor: Two radically different evolutionary patterns within green algae. Plant Cell 11, 1717–1729 (1999).
30. Boyen, C., Leblanc, C., Bonnard, G., Grienengerber, J. M. & Kloareg, B. Nucleotide-Sequence of the Cox3 Gene from Chondrus crispus - Evidence That Uga Encodes Tryptophan and Evolutionary Implications. Nucleic Acids Res 22, 1400–1403 (1994).
31. Macino, G., Coruzzi, G., Nobrega, F. G., Li, M. & Tzagoloff, A. Of the Uga Terminator as a Tryptophan Codon in Yeast Mitochondria. P Natl Acad Sci USA 76, 3784–3785 (1979).
32. Beagley, C. T., Okimoto, R. & Wolstenholme, D. R. The mitochondrial genome of the sea anemone Metridium senile (Cnidaria): Introns, a paucity of tRNA genes, and a near-standard genetic code. Genetics 148, 1091–1108 (1998).
33. Bessho, Y., Ohama, T. & Osawa, S. Planarian Mitochondria. 2. The Unique Genetic Code as Deduced from Cytochrome-C-Oxidase Subunits I Gene- Sequences. J Mol Evol 34, 331–335 (1992).
34. Telford, M. J., Herniou, E. A. & Russell, R. R., Littlewood DTJ. Changes in mitochondrial genetic codes as phylogenetic characters: Two examples from the flatworms. P Natl Acad Sci USA 97, 11359–11364 (2000).
35. Hoffmann, R. I., Boone, J. L. & Brown, W. M. A novel mitochondrial genome organization for the blue mussel, Mytilus edulis. Genetics 131, 597–412 (1992).
36. Jacobs, H. T., Elliott, D. J., Math, V. B. & Farquharson, A. Nucleotide sequence and gene organization of sea urchin mitochondrial DNA. J Mol Biol 202, 185–217 (1988).
37. Boone, I. L., Dachler, L. L. & Brown, W. M. Complete sequence, gene arrangement, and genetic code of mitochondrial DNA of the cephalochordate Branchiostoma floridae (Amphioxus). Mol Biol Evol 16, 410–418 (1999).
38. Barrell, B. G., Bankier, A. T. & Drouin, J. A different genetic code in human mitochondria. Nature 282, 189–194 (1979).
39. Clark, G. D. & Welller, G. F. The Structure of the Small Mitochondrial-DNA of Kluiveromyces Thermotolerans Is Likely to Reflect the Ancestral Gene Order in Fungi. J Mol Evol 38, 593–601 (1994).
40. Ebara, M., Hayashishimaru, Y., Inagaki, Y. & Ohama, T. Use of a deviant mitochondrial genetic code in yellow-green algae as a landmark for segregating members within the phylum. J Mol Evol 45, 119–124 (1997).
41. Kruit, V., Eubel, H., Jansch, L., Werhahn, W. & Braun, H. P. Proteomic approach to identify novel mitochondrial proteins in Arabidopsis. Plant Physiol 127, 1694–1710 (2001).
42. Schultz, D. W., Yarun, M. & Transfer, R. N. A. mutation and the malleability of the genetic code. J Mol Biol 235, 1377–1380 (1994).
43. Schultz, D. W. & Yarun, M. On malleability of the genetic code. J Mol Evol 42, 597–601 (1996).
44. Sengupta, S. & Higgs, P. G. The unified model of codon reassignment in alternative genetic codes. Genetics 170, 831–840 (2005).
45. Seaborg, D. M. Was Wright right? The canonical genetic code is an empirical example of an adaptive peak in nature; deviant genetic codes evolved using adaptive bridges. J Mol Evol 71, 87–99 (2010).
46. Pauling, L., Zucker, R., Hendriksen, E. & Lövstad, R. Chemical Paleogenetics. Molecular “Restoration Studies” of Extinct Forms of Life. Acta Chemica Scandinavica 17 suppl, 9–16 (1963).
47. Maddison, W. P. Calculating the Probability Distributions of Ancestral States Reconstructed by Parsimony on Phylogenetic Trees. *Systematic Biology* **44**, 474–481 (1995).
48. Schultz Gact, R. The Role of Subjectivity in Reconstructing Ancestral Character States: A Bayesian Approach to Unknown Rates, States, and Transformation Asymmetries. *Systematic Biology* **48**, 651–664 (1999).
49. Yang, Z. & Roberts, D. On the use of nucleic acid sequences to infer early branchings in the tree of life. *Mol Biol Evol* **12**, 451–458 (1995).
50. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792–1797 (2004).
51. Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113 (2004).
52. Hedges, S. B., Blair, J. E., Venturi, M. L. & Shoe, J. L. A molecular timescale of eukaryote evolution and the rise of complex multicellular life. *BMC evolutionary biology* **4**, 2 (2004).
53. Heckman, D. S., Geiser, D. M., Eidell, B. R., Stauffer, R. L., Kardos, N. L. & Hedges, S. B. Molecular evidence for the early colonization of land by fungi and plants. *Science* **293**, 1129–1133 (2001).
54. Sanderson, M. J., Thorne, J. L., Wikstrom, N. & Bremer, K. Molecular evidence on plant divergence times. *Am J Bot* **91**, 1656–1665 (2004).
55. Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC evolutionary biology* **7**, 214 (2007).

**Acknowledgments**

This work was supported by São Paulo Research Foundation (FAPESP) grants 2008/52067-3 and 2008/04992-0 to MCSF, 2011/00417-3 provided to MMB. This work was also supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) grants 303059/2010-9 and 503891/2011-8 provided to RPJ. MCSF and RPJ are also research fellows at CNPq.

**Author Contributions**

M.M.B. conducted all of the phylogenetic and computational biology analyses. L.S. performed the codon reassignment search and analyses. L.C.B.F., A.S.L.R. and R.P.J. developed the DNA-SGA, L.C.B.F. and A.S.L.R. assembled the code-generated sequence database. M.M.B., M.C.S.F. and R.P.J. proposed the main concept of this manuscript, and M.C.S.F. and R.P.J. jointly supervised the work. L.S., L.C.B.F. and A.S.L.R. contributed equally to the work.

**Additional Information**

**Supplementary information** accompanies this paper at http://www.nature.com/srep

**Competing financial interests:** The authors declare no competing financial interests.

**How to cite this article:** Brandão, M. M. et al. Ancient DNA sequence revealed by error-correcting codes. *Sci. Rep.* **5**, 12051; doi: 10.1038/srep12051 (2015).

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/