# Genomic Abelian Finite Groups

Robersy Sanchez [1] and Jesús Barreto[2]

[1]Department of Biology. Pennsylvania State University, University Park, PA 16802.
E-mail: rus547@psu.edu
ORCID: https://orcid.org/0000-0002-5246-1453

[2]Universidad Central "Marta Abreu" de Las Villas. Santa Clara. Cuba.
E-mail: barretouclv@gmail.com

[1] Corresponding author:
rus547@psu.edu

## Abstract

Experimental studies reveal that genome architecture splits into natural domains suggesting a well-structured genomic architecture, where, for each species, genome populations are integrated by individual mutational variants. Herein, we show that the architecture of population genomes from the same or closed related species can be quantitatively represented in terms of the direct sum of homocyclic abelian groups defined on the genetic code, where populations from the same species lead to the same canonical decomposition into *p*-groups. This finding unveils a new ground for the application of the abelian group theory to genomics and epigenomics, opening new horizons for the study of the biological processes (at genomic scale) and provides new lens for genomic medicine.

**Keywords**: Genomics, Genetic code, Abelian groups, genome algebra

# 1 Introduction

The analysis of the genome architecture is one of biggest challenges for the current and future genomics. Current bioinformatic tools make possible faster genome annotation process than some years ago [2]. Current experimental genomic studies suggest that genome architectures must obey specific mathematical biophysics rules [3–6].

Experimental results points to an injective relationship: *DNA sequence → 3D chromatin architecture* [3,4,6], and failures of DNA repair mechanisms in preserving the integrity of the DNA sequences lead to dysfunctional genomic rearrangements which frequently are reported in several diseases [5]. Hence, some hierarchical logic is inherent to the genetic information system that makes it feasible for mathematical studies. In particular, there exist mathematical biology reasons to analyze the genetic information system as a communication system [7–10].

Under the assumption that current forms of life evolved from simple primordial cells with very simple genomic structure and robust coding apparatus, the genetic code is a fundamental link to the primeval form of live, which played an essential role on the primordial architecture. The genetic code, the set of biochemical rules used by living cells to translate information encoded within genetic material into proteins, sets the basis for our understanding of the mathematical logic inherent to the genetic information system [9,11]. The genetic code is the cornerstone of live on earth. Not a single form of live could evolve or exist, as we currently know it, without the genetic code.

Several genetic code algebraic structures has been introduced to study effect of the quantitative relationship between the coding apparatus and the mutational process on protein-coding regions [12–16]. Formally the genetic code only is limited to translated coding regions where the number of RNA bases is a multiple of 3. However, as suggested in reference [17], the difficulties in prebiotic synthesis of the nucleosides components of RNA (nucleo-base + sugar) and suggested that some of the original bases may not have been the present purines or pyrimidines [18]. Piccirilli et al. [19] demonstrated that the alphabet can in principle be larger. Switzer et al. [20] have shown an enzymatic incorporation of new functionalized bases into RNA and DNA. This expanded the genetic alphabet from 4 to 5 or more letters, which permits new base pairs, and provides RNA molecules with the potential to greatly increase their catalytic power.

It is important to notice that even in the current (*friendly*) environmental conditions not a single cell can survive without a DNA repair enzymatic machinery and that such an enzymatic machinery did not existed at all in the primaeval forms of live. Here, we are confronting the *chicken and egg* problem. To date, the best solution (to our knowledge) is the admission of alternative base-pairs in the primordial DNA alphabet which, as suggested in the studies on the prebiotic chemistry, could contribute to the thermal and general physicochemical stability of the primordial DNA molecules.

Several algebraic structures have been proposed including an additional letter into the DNA alphabet: A, C, G, T. The new letter (D) stands for current insertion deletion/mutations or for alternative wobble base pairing, which would be a relict fingerprint from primordial enzymes derived from a more degenerated ancestral genetic code [17,21,22]. Supporting evidence for the existence of a more degenerated ancestral genetic code built up on a larger alphabet is found in the tRNA anticodon region permitting wobble base pairing by including, e.g., bases such as: inosine (in eukariotes), agmatidine (in archaea), and lysidine (in bacteria), which has been proposed as evolutionary solutions to the need for lower the high translational noise connected to the reading of the AUA and AUG codons [23,24]. Additionally, various alternative base pairs like methylated cytosine and adenine are still present in the current genomes playing an important role in the epigenetic adaptation of organismal populations to the continuous environmental changes [10,25].

Cytosine DNA methylation results from the addition of methyl groups to cytosine C5 residues, and the configuration of methylation within a genome provides trans-generational epigenetic information. These epigenetic modifications can influence the transcriptional activity of the corresponding genes, or maintain genome integrity by repressing transposable elements and affecting long-term gene silencing mechanisms [26,27].

In this scenario, we shall show that all possible DNA molecules and, consequently, genomes can be described by way of finite abelian groups which can be split into the direct sum of homocyclic 2-groups and 5-groups defined on the genetic code. A homocyclic group is a direct sum of cyclic groups of the same order. Any finite abelian group can be decomposed into a direct sum of homocyclic $p$-groups [28], i.e., a group in which the order of every element is a power of a primer number $p$.

The genetic code algebraic structures under scrutiny in the mentioned references covered rings and vector spaces with a common feature, the corresponding additive group is an abelian group of prime-power order. Next, to help a better comprehension of the current work, a brief introductory summary on these groups is provided. Results presented here generalizes the application of the genetic code algebras (reported in several publications) to the whole genome.

## 1.1 Reported genetic code abelian groups relevant for the current study

Herein, we assume that readers are familiar with the definition of abelian group, which otherwise can be found in textbooks and elsewhere including Wikipedia. Nevertheless, all the abelian groups discussed here are isomorphic to the well-known abelian groups of integer module $n$, which are easily apprehended by a college-average educated mind. For example, the abelian group defined on the set {0, 1, 2, 3, 4}, which corresponds to the group of integer modulo 5 ( $\mathbb{Z}_5$), where (2 + 1) mod 5 = 3, (1 + 3) mod 5 = 4, (2 + 3) mod 5 = 0, etc. The subjacent biophysical and biochemical reasonings to define the algebraic operations on the set of DNA bases and on the codon set were given in references [12,14,17].

### *1.1.1 The $\mathbb{Z}_{64}-$algebras of the genetic code ($C_g$)*

The $\mathbb{Z}_{64}-$algebras of the genetic code ($C_g$) and gene sequences were stated several years ago. In the $\mathbb{Z}_{64}-$algebra $C_g$ the sum operation, defined on the codon set, is a manner to consecutively obtain all codons from the codon AAC (UUG) in such a way that the genetic code will represent a non-dimensional code scale of amino acids interaction energy in proteins.

A description of the genetic code abelian finite group ($C_g$, +) can be found in [12]. Group $\left(C_g,+\right)$ is isomorphic to the group on the set $\mathbb{Z}_{2^6}$ (the sum of integer modulo 64), which formally will be expressed as $\left(C_g,+\right)\cong\left(\mathbb{Z}_{2^6},+\right)$. This group on the set $\mathbb{Z}_{2^6}$ (the sum of integer modulo 64).

The mapping of the set of codons $X_1X_2X_3 \in C_g$ into the set $\mathbb{Z}_{2^6}$ is straightforward after consider the bijection $\mathrm{A} \leftrightarrow 0, \mathrm{C} \leftrightarrow 1, \mathrm{G} \leftrightarrow 2, \mathrm{U} \leftrightarrow 3$ and the function $g(x) = 4x_1 + 16x_2 + x_3$. For example:

$$\begin{array}{rcr} \mathrm{AGC} & \leftrightarrow & 33 \\ +\,\mathrm{UGU} & \leftrightarrow & +47 \\ \hline \mathrm{ACA} & \leftrightarrow & 16 \bmod 64 \end{array} \qquad \begin{array}{rcr} \mathrm{AGC} & \leftrightarrow & 33 \\ +\mathrm{ACU} & \leftrightarrow & +18 \\ \hline \mathrm{AUU} & \leftrightarrow & 51 \bmod 64 \end{array} \qquad \begin{array}{rcr} \mathrm{GGC} & \leftrightarrow & 41 \\ \mathrm{CUA} & \leftrightarrow & +52 \\ \hline \mathrm{UCC} & \leftrightarrow & 29 \bmod 64 \end{array}$$

The $Z_{64}$-algebra $C_g$, however, is limited to protein-coding regions, while it is well known that, in eukaryotes, only a small fraction of the genome –about 3%– called open reading frame (ORF) encodes for proteins [18]. Since non-coding DNA sequences can have a base pairs number not multiple of three, complete chromosomes and genomes cannot be described by means of group $(C_g, +)$. In addition, natural genomic variations that includes insertions and deletion mutations (indel mutations) across individuals from the same population and close-related populations from different species cannot be represented with group $(C_g, +)$.

### *1.1.2 The $(\mathbb{Z}_2^6, +)$ group of the genetic code ($C_g$)*

Group $(C_g, +)$ is the additive group of a module over a ring, which however, do not conform to a vector space. To build a genetic code vector space, a Galois field (*GF*(4) structure in the ordered base set $B = \{\mathrm{G}, \mathrm{U}, \mathrm{A}, \mathrm{C}\}$ was introduced in reference [14]. In particular, an isomorphism with the Galois field is defined by means of its binary representation $\mathbb{Z}_2 \times \mathbb{Z}_2$= {(0, 0), (0, 1), (1, 0), (1, 1)}, i.e. a unique *GF*(4) up to isomorphism exists, such that a bijection $f : \{\mathrm{G,U,A,C}\} \leftrightarrow \mathbb{Z}_2 \times \mathbb{Z}_2$ from the DNA base set $B$ = {G, U, A, C} to the set of binary duplets ($\alpha_1$, $\alpha_2$) is stated., where $\alpha_i \in \mathbb{Z}_2 = \{0,1\}$ , for $i \in \{1, 2\}$. For example, the bijection *f* is defined as:

$$f(\mathrm{G}) = (0, 0), f(\mathrm{U}) = (0, 1), f(\mathrm{A}) = (1, 0), f(\mathrm{C}) = (1, 1).$$

The additive group of bases is the Klein four-group, which is defined by the group presentation: $V = \{\mathrm{U,A} \mid \mathrm{U+U=A+A=C+C=G,\ A+U=C}\}$, i.e., $(B, +) \cong (\mathbb{Z}_2^2, +)$. Next, the abelian group on the set of codons $B^3$ was defined as the direct third power $B^3 = B \times B \times B$ of the group ($B$, +), i.e.

$\left(B^3,+\right)=\left(B,+\right)\times\left(B,+\right)\times\left(B,+\right)$, which is isomorphic to the group: $\left(\mathbb{Z}_2^6,+\right)=\left(\mathbb{Z}_2^2,+\right)\times\left(\mathbb{Z}_2^2,+\right)\times\left(\mathbb{Z}_2^2,+\right)$ , i.e., $\left(B^3,+\right)\cong\left(\mathbb{Z}_2^6,+\right)$. The sum operation on the set $\left(B^3,+\right)$ follows from the sum operation by coordinates.

As pointed out before by Crick, the first two bases of codons determine the physicochemical properties of aminoacids [29]. The four encoded amino acids of every class are either the same or show very similar physicochemical properties. This genetic code regularity is captured by the quotient group $B^3/G_{\mathrm{GGA}}$ , where $G_{\mathrm{GGA}}$ is a subgroup of $B^3$ integrated by the elements {GGG,GGA} (the elements of the quotient group $B^3/G_{\mathrm{GGA}}$ are given in Table 5 from [14]). The quotient group $B^3/G_{\mathrm{GGA}}$ is isomorphic to group $\left(\mathbb{Z}_2^5,+\right)=\left(\mathbb{Z}_2^2,+\right)\times\left(\mathbb{Z}_2^2,+\right)\times\left(\mathbb{Z}_2,+\right)$. Each element of this group represents an equivalence class of codons. Two triplets $X_1X_2X_3$ and $Y_1Y_2Y_3$ are equivalent if, and only if, the difference $X_1X_2X_3+Y_1Y_2Y_3\in G_{DD\mathrm{A}}$. In biological terms, substitution mutations involving codons from the same class will not alter (or at least no substantially alter in most of the cases) the physicochemical properties of the encoded protein domains, since in the worst scenario involves aminoacids with very close physicochemical properties, with the exception of codon for aminoacid tryptophan.

### *1.1.3 The $\mathbb{Z}_{125}$ group of the extended genetic code ($C_e$)*

The extension of the *genetic code group* $\left(C_g,+\right)$ follows straightforward from the extension of the codon set, which is easily accomplished extending the source alphabet of the standard genetic code: {A, C, G, U} and, consequently, extending the base triplet set (extended triplet) as $X_1X_2X_3$, $X_i\in${D, A, C, G, U} [22]. The new algebraic structure $\left(C_e,+\right)$ is isomorphic to the abelian group defined on the set $\mathbb{Z}_{5^3}$ (the sum of integer modulo 125), formally, $\left(C_e,+\right)\cong\left(\mathbb{Z}_{5^3},+\right)$. The mapping of the set of codons $X_1X_2X_3\in C_e$ into the set $\mathbb{Z}_{5^3}$ is straightforward after consider the bijection

$D \leftrightarrow 0, \text{A} \leftrightarrow 1, \text{C} \leftrightarrow 2, \text{G} \leftrightarrow 3, \text{U} \leftrightarrow 4$ and the function $g(x) = 5x_1 + 25x_2 + x_3$ (see Table 1). For example:

$$\begin{array}{rcr} \text{AGC} & \leftrightarrow & 82 \\ +\,\text{UGU} & \leftrightarrow & +99 \\ \hline \text{ACA} & \leftrightarrow & 56 \bmod 125 \end{array} \qquad \begin{array}{rcr} \text{AGC} & \leftrightarrow & 82 \\ +\,\text{DCU} & \leftrightarrow & +54 \\ \hline \text{CDA} & \leftrightarrow & 11 \bmod 125 \end{array} \qquad \begin{array}{rcr} \text{GGC} & \leftrightarrow & 92 \\ +\,\text{CUD} & \leftrightarrow & +110 \\ \hline \text{DGC} & \leftrightarrow & 77 \bmod 125 \end{array}$$

**Table 1**. Ordered set of extended triplets corresponding to the elements from $\mathbb{Z}_{5^3}$

| | D | | A | | | C | | | G | | | U | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| a | I | III | I | III | | I | III | | I | III | | I | III | | |
| | 0 | DDD | 25 | DAD | | 50 | DCD | | 75 | DGD | | 100 | DUD | | D |
| | 1 | DDA | 26 | DAA | | 51 | DCA | | 76 | DGA | | 101 | DUA | | A |
| D | 2 | DDC | 27 | DAC | | 52 | DCC | | 77 | DGC | | 102 | DUC | | C |
| | 3 | DDG | 28 | DAG | | 53 | DCG | | 78 | DGG | | 103 | DUG | | G |
| | 4 | DDU | 29 | DAU | | 54 | DCU | | 79 | DGU | | 104 | DUU | | U |
| | 5 | ADD | 30 | AAD | | 55 | ACD | | 80 | AGD | | 105 | AUD | | D |
| | 6 | ADA | 31 | AAA | K | 56 | ACA | T | 81 | AGA | R | 106 | AUA | I | A |
| A | 7 | ADC | 32 | AAC | N | 57 | ACC | T | 82 | AGC | S | 107 | AUC | I | C |
| | 8 | ADG | 33 | AAG | K | 58 | ACG | T | 83 | AGG | R | 108 | AUG | M | G |
| | 9 | ADU | 34 | AAU | N | 59 | ACU | T | 84 | AGU | S | 109 | AUU | I | U |
| | 10 | CDD | 35 | CAD | | 60 | CCD | | 85 | CGD | | 110 | CUD | | D |
| | 11 | CDA | 36 | CAA | Q | 61 | CCA | P | 86 | CGA | R | 111 | CUA | L | A |
| C | 12 | CDC | 37 | CAC | H | 62 | CCC | P | 87 | CGC | R | 112 | CUC | L | C |
| | 13 | CDG | 38 | CAG | Q | 63 | CCG | P | 88 | CGG | R | 113 | CUG | L | G |
| | 14 | CDU | 39 | CAU | H | 64 | CCU | P | 89 | CGU | R | 114 | CUU | L | U |
| | 15 | GDD | 40 | GAD | | 65 | GCD | | 90 | GGD | | 115 | GUD | | D |
| | 16 | GDA | 41 | GAA | E | 66 | GCA | A | 91 | GGA | G | 116 | GUA | V | A |
| G | 17 | GDC | 42 | GAC | D | 67 | GCC | A | 92 | GGC | G | 117 | GUC | V | C |
| | 18 | GDG | 43 | GAG | E | 68 | GCG | A | 93 | GGG | G | 118 | GUG | V | G |
| | 19 | GDU | 44 | GAU | D | 69 | GCU | A | 94 | GGU | G | 119 | GUU | V | U |
| | 20 | UDD | 45 | UAD | | 70 | UCD | | 95 | UGD | | 120 | UUD | | D |
| | 21 | UDA | 46 | UAA | Stop | 71 | UCA | S | 96 | UGA | Stop | 121 | UUA | L | A |
| U | 22 | UDC | 47 | UAC | Y | 72 | UCC | S | 97 | UGC | C | 122 | UUC | F | C |
| | 23 | UDG | 48 | UAG | Stop | 73 | UCG | S | 98 | UGG | W | 123 | UUG | L | G |
| | 24 | UDU | 49 | UAU | Y | 74 | UCU | S | 99 | UGU | C | 124 | UUU | F | U |

[a] Bijection between the base-triplets set and the elements from sets $\mathbb{Z}_{5^3}$ as given in [22].

### *1.1.4 The $(\mathbb{Z}_5^3, +)$ group of the extended genetic code ($C_e$)*

The Galois field *GF*(5) of the DNA set of bases $\mathfrak{B} = \{\text{D}, \text{A}, \text{C}, \text{G}, \text{U}\}$ was introduced in reference [17]. This structure led to the definition of a $\mathbb{Z}_5$ −vector space $\mathfrak{B}^3$ over the set $\mathfrak{B}^3 = \mathfrak{B} \times \mathfrak{B} \times \mathfrak{B}$ isomorphic to the set $\mathbb{Z}_5^3 = \mathbb{Z}_5 \times \mathbb{Z}_5 \times \mathbb{Z}_5$ [17,30]. But here, we are interested only in the abelian groups $(\mathfrak{B}, +)$ and $(\mathfrak{B}^3, +)$. After the bijection $D \leftrightarrow 0, \text{A} \leftrightarrow 1, \text{C} \leftrightarrow 2, \text{G} \leftrightarrow 3, \text{U} \leftrightarrow 4$, the sum operation of two DNA bases follows from the sum operation on the Galois field *GF*(5) (i.e., on $\mathbb{Z}_5$, the sum of integers modulo 5). For example, $\mathrm{C}+\mathrm{U} \leftrightarrow (2+4) \bmod 5 = 1 \leftrightarrow \mathrm{A}$. The sum operation on the set $\mathfrak{B}^3$ follows from the sum operation by coordinates.

It is worthy to notice that there 24 way to define each one of the above mentioned algebraic structures [30,31]. Nevertheless, for each defined genetic code group, there is only one (genetic code abelian group) up to isomorphism, which lead to their representation as an abelian group, where the sum operation corresponds to the sum of integer modulo $n \in \{2, 2^6, 5, 5^3\}$.

## 2 The General Theoretical Model

Herein, it will be showed that, in a general scenario, the whole genome population from any species or close related species, can be algebraically represented as a direct sum of abelian cyclic groups or more specifically abelian *p*-groups. Basically, we propose the representation of multiple sequence alignments (MSA) of length *N* as the direct sum:

$$G = \left(\mathbb{Z}_{p_1}\right)^{n_1} \oplus \left(\mathbb{Z}_{p_2}\right)^{n_2} \oplus \cdots \oplus \left(\mathbb{Z}_{p_k}\right)^{n_k} \qquad [1]$$

Where $p_i \in \{2, 5, 2^6, 5^3\}$ and $N = n_1 + n_2 + ... + n_k$. Here, we assume the usual definition of direct sum of groups [32]. Let $B_i$ $(i \in I = \{1, ..., n\})$ be a family of subgroups of $G$, subject to the following two conditions:

1) $\sum B_i = G$. That is, $B_i$ together generates $G$.
2) For every $i \in I$: $B_i \cap \sum B_j = 0$.

Then, it is said that $G$ is the direct sum of its subgroups $B_i$, which formally is expressed by the expression: $G = \underset{i}{\oplus} B_i$ or $G = B_1 \oplus \ldots \oplus B_n$.

In superior organisms, genomic DNA sequences are integrated by intergenic regions and gene regions. The former are the larger regions, while the later includes the protein-coding regions as subsets. The MSA of DNA and protein-coding sequences reveals allocations of the nucleotide bases and aminoacids into stretched of *strings*. The alignment of these stretched would indicate the presence of substitution, *indel* mutations. As a result, the alignment of a whole chromosome DNA sequences from several individuals from the same or close-related species can be split into well-defined subregions or domains, and each one of them can be represented as homocyclic abelian groups, i.e., a cyclic group of *prime-power* order (Fig. 1). As a result, each DNA sequence is represented as a *N*-dimensional vector with numerical coordinates representing bases and codons.

**Fig.1**. A typical DNA multiple sequence alignment (MSA) including segments of protein-coding regions. A MSA would include the presence of substitution, insertion and deletion mutations (*indel* mutations). The aligned sequences can be grouped into blocks, which can be algebraically represented by abelian groups.

An intuitive mathematical representation of MSA is implicit in Fig.1, with following observations:

a) Every DNA sequence from the MSA and every subsequence on it can be represented as a vector with element coordinates defined in some abelian group. For example, $\left(C_g,+\right)\cong\left(\mathbb{Z}_{64},+\right)$, the first five codons from the first DNA sequence from Fig. 1, $\left\{\text{ATA, CCC, ATG, GCC, AAC}\right\}\in\left(C_g,+\right)$, can be represented by the vector of integers: $\left\{48,21,50,25,1\right\}$ where each coordinate is an element from group $\left(\mathbb{Z}_{64},+\right)$ (see Table 1 from reference [12] and the introduction section).

b) Any MSA can be algebraically represented as a symbolic composition of abelian groups each one of them is isomorphic to an abelian group of integers module *n*. Such a composition can be algebraically represented as a direct sum of homocyclic abelian groups. For example, the multiple sequence alignment from Fig. 1 can be represented by the direct sum of abelian groups:

$$G=(\mathbb{Z}_{2^6})^5\oplus(\mathbb{Z}_5)^8\oplus(\mathbb{Z}_{5^3})^5\oplus(\mathbb{Z}_5)^7\oplus(\mathbb{Z}_{5^3})^4 \qquad [2]$$

In more specific scenario, the multiple sequence alignment from Fig. 1 can be represented by the direct sum of abelian 2-groups and 5-groups:

$$G = (\mathbb{Z}_2^6)^5 \oplus (\mathbb{Z}_5)^8 \oplus (\mathbb{Z}_{5^3})^5 \oplus (\mathbb{Z}_5)^7 \oplus (\mathbb{Z}_2^6)^4 \qquad [3]$$

Or strictly as the direct sum of abelian 5-groups:

$$G = (\mathbb{Z}_5^3)^5 \oplus (\mathbb{Z}_5)^8 \oplus (\mathbb{Z}_{5^3})^5 \oplus (\mathbb{Z}_5)^7 \oplus (\mathbb{Z}_5^3)^4 \qquad [4]$$

Although the above *direct sums* of abelian groups provides a useful compact representation of MSA, for application purposes to genomics, we would also consider to use the concept of direct product (*cartesian sum or complete direct sums*) [32]. Next, let *S* be a set of abelian cyclic groups identified in the MSA *M* of length *N* (i.e., every DNA sequence from *M* has *N* bases). Let $\ell_i$ the number of bases or triples of bases covered on *M* by group $S_i \in S$ where $\sum_i \ell_i = N$. Hence, each DNA sequence on the *M* can be represented by a cartesian product $(b_1,\dots,b_n)$ where $b_i \in S_i$ $(i=1,\dots,n)$ and $n=|S|$. Let $\mathcal{G}_i$ be a group defined on the set of all elements $(0,\dots,0,b_i,0,\dots0)$ where $b_i \in S_i$ stands on the $i^{th}$ place and 0 everywhere else. It is clear that $S_i \cong \mathcal{G}_i$. In this context, the set of all vectors $(b_1,\dots,b_n)$ with equality and addition of vectors defined coordinate-wise becomes a group ($\mathcal{G}$) named direct product (cartesian sum) of groups $S_i$ ($\mathcal{G}_i$), i.e.:

$$\mathcal{G} = \otimes_i S_i = \oplus_i \mathcal{G}_i \qquad [5]$$

An illustration of the cartesian sum application was given above in observation a).

## 3 Results

Results essentially comprise an application of the fundamental theorem of abelian finite groups [28,32]. By this theorem every finite abelian group *G* is isomorphic to a direct sum of cyclic groups of prime-power order of the form:

$$G = \mathbb{Z}_{p_1^{\alpha_1}} \oplus \mathbb{Z}_{p_2^{\alpha_2}} \oplus \cdots \oplus \mathbb{Z}_{p_n^{\alpha_n}} \qquad [6]$$

Or (in short) $G = \oplus_{i=1}^n \mathbb{Z}_{p_i^{\alpha_i}}$, where the $p_i$'s are primes (not necessarily distinct), $\alpha_i \in \mathbb{N}$ and $\mathbb{Z}_{p_i^{\alpha_i}}$ is the group of integer module $p_i^{\alpha_i}$. The abelian group representation of the MSA from Fig. 1 given by expressions [1] and [2] correspond to the cases where the finite abelian group *G* is a direct sum of *prime-power order*, while expression [3] reflects the fact that any finite abelian group can be decomposed into a direct sum of homocyclic $p$-groups [28,32], in this $p = 5$.

As is showed in Fig 1, this abelian group is a heterocyclic group that split into a direct sum of homocyclic *prime-power order*, each one of them split into the direct sum of cyclic $p$-groups with same order. For example, in expression [4] we have the subgroup: $\left(\mathbb{Z}_5^3\right)^4 = \oplus_{i=1}^{12}\mathbb{Z}_5$, which is a direct sum of 12 homocyclic 5-groups $\left(\mathbb{Z}_5,+\right) \cong \left(\mathfrak{B},+\right)$. The case of $\mathbb{Z}_{2^6}$ representation of the genetic code (as given in [12]) is less evident. It follows from the fact that the genetic code table is integrated by 16 subsets of codons with form $K = \left\{XY\mathrm{A}, XY\mathrm{C}, XY\mathrm{G}, XY\mathrm{U}\right\}$, where $X \in B$ and $Y \in B$ are fixed, the sum operation on each set $K$ is defined by coordinates as in the set of bases $\left(B,\otimes\right)$, and codon $XY$A is taken as identity element. For example, $K$= {CGA, CGC, CGG, CGU} with codon CGA as identity element. In other words, $\left(K,+\right) \cong \left(B,\otimes\right) \cong \left(\mathbb{Z}_2^2,+\right)$, which corresponds to the Klein four group as defined on $\mathbb{Z}_2^2$.

Notice that for each fixed length $N$ we can build manifold heterocyclic groups $S_i$, and each one of them can have different decomposition into $p$-groups. So, each group $S_i$ could be characterized by means of their corresponding canonical decomposition into $p$-groups. This last detail is exemplified in Fig. 2, where an exon region from the enzyme *phospholipase B domain containing-2* (PLBD2) simultaneously encodes information for several aminoacids and carries the footprint to be targeted by the transcription factor REST. Four possible group representations for this exon subregion are suggested in the top of the figure (panel **a**). These types of protein-coding regions are called *duons*, since their base-triplets encode information not only for aminoacids but also for transcription enhancers [33–35].

**Fig. 2**. The DNA sequence motifs targeted by transcription factors usually integrate genomic building block across several species. **a**, DNA sequence alignment of the protein-coding sequences from phospholipase B domain containing-2 (PLBD2) carrying the footprint sequence motif recognized (targeted) by the Silencing Transcription factor (REST), also known as Neuron-Restrictive Silencer Factor (NRSF) REST (NRSF). **b**. Sequence logo of the footprint motif recognized REST (NRSF) on the exons (derived from TF2DNA dataset [36]). **c**, Translation of the codon sequences using the one-letter symbol of the aminoacids.

The group representation is particularly interesting for the analysis of DNA sequence motifs, which typically are highly conserved across the species. As suggested in Fig. 2, there are some subregions of DNA or protein sequences where there are few or not gaps introduced and mostly substitution mutations are found. Such subregions conform blocks that can cover complete DNA sequence motifs targeted by DNA biding proteins like transcription factors, which are identifiable by bioinformatic algorithm like BLAST [37]. Herein, the case of double coding called our attention, where the DNA sequence simultaneously encode information transcription factor targeted sequence motif and the codon sequence encoding for aminoacids. Notice that the abelian group $\left(C_g,+\right)\cong\left(\mathbb{Z}_{64},+\right)$ defined on the standard genetic code is enough to quantitatively describe these motifs (Fig. 2). However, a further application of group theory together with additional knowledge on the biological function this motif can lead to a more specific decomposition into abelian groups.

No matter how complex a genomic region might be, it has an abelian group representation. As shown in Fig. 3, two different protein-coding (gene) models from two different genome populations can lead to the same direct sum of abelian $p$-groups and the same final aminoacids sequence (protein). The respective exon regions have different lengths and gaps (“-”, representing base D in the extended genetic code) were added to exons 1 and 2 (from panel **a**) to preserve the reading frame in the group representation (after transcription and splicing gaps are removed). Both gene models, from panel **a** and **b**, however, lead to the same direct sum of abelian *5*-groups:

$$\left(\mathbb{Z}_5\right)^{n+7}\oplus\left(\mathbb{Z}_5^3\right)^3\oplus\left(\mathbb{Z}_5\right)^{3+m+m+2}\oplus\left(\mathbb{Z}_5^3\right)^3\oplus\left(\mathbb{Z}_5\right)^{n+8}.$$

**Fig. 3**. Two different protein-coding (gene) models can lead to the same abelian group representation and the same protein sequence. A dummy intron was drawn carrying the typical sequence motif targeted by the spliceosome the donor (GU$R$) and acceptor ($Y^m$AG) sites, where $R\in\{\mathrm{A},\mathrm{G}\}$ (purines) and $Y\in\{\mathrm{C},\mathrm{U}\}$, $X$ stands for any base, and $n$ and $m$ indicate the number of bases present in the corresponding sub-sequences (pyrimidines). **a**, A gene model based on a *dummy* consensus sequence where gaps representing base D from the extended genetic code were added to preserve the coding frame, which naturally is restored by splicing soon after transcription. **b**, A gene model where both exons, 1 and 2, carries a complete set of three codons (base-triplets). Both models, from panels **a** and **b**, leads to the same canonical direct sum of abelian 5-groups.

An example considering changes on the gene-body reading frames as those introduced by alternative splicing is shown in Fig. 4. Gene-bodies with annotated alternative splicing can easily be represented by any of the groups $(\mathbb{Z}_{5^3})^n$or $\left(\mathbb{Z}_5^3\right)^n$ (Fig.4a). The splicing scenario can include enhancer regions as well (Fig.4b).

**Fig. 4**. The abelian group representation of a given genome only depend on our current knowledge on its annotation. **a,** the alternative splicing specified for an annotated gene model does not alter the abelian group representation and only would add information for the decomposition of the existing cyclic groups into subgroups. **b**, a more complex gene model including detailed information on the promoter regions. A GC box (G5MG4CU) motif is located upstream of a TATA box (TATAWAW) motif in the promoter region. The GC box is commonly the binding site for Zinc finger proteins, particularly, Sp1 transcription factors. A putative GC box was included in exon 2, which is an atypical scenario, but it can be found, e.g., in the second exon from the gene encoding for sphingosine kinase 1 (SPHK1), transcript variant 2 (NM_182965, CCDS11744.1).

As commented in the introduction, cytosine DNA methylation is implicitly included in extended base-triple group representation. Typically, methylation analysis of methylomes is addressed to identify methylation changes induced by, for example, environmental changes, lifestyles, age, or diseases. So, in this case the letter D stands for methylated cytosine, since only epigenetic changes are evaluated. A concrete example with two genes from patients with pediatric acute lymphoblastic leukemia (PALL) is presented in Fig. 5.

**Fig. 5**. Vector representation of differentially methylated exons regions from genes EGEL7 and P2RY1 from patients with pediatric acute lymphoblastic leukemia (PALL). **a**. Segment of exon-6 from gene EGFL7. **b**. Segment of exon-1 from gene P2RY1. Methylated cytosines are highlighted in yellow background. In PALL patients, gene EGEL7 mostly hypomethylated and gene P2RY1 mostly hypermethylated in respect to healthy individuals (WT). The encoded aminoacid sequence is given using the one letter symbols. Both genes, EGEL7 and P2RY1, were identified in the top ranked list of differentially methylated genes integrating clusters of hubs in the protein-protein interaction networks from PALL reported in reference [38].

It is obvious that the MSA from a whole genome population derives from the MSA of every genomic region, from the same or closed related species. At this point, it is worthy to recall that there is not, for example, just one human genome or just one from any other species, but populations of human genomes and genomes populations from other species. Since every genomic region can be represented by the direct sum of abelian cyclic groups of prime-power order, then the whole genome population from individuals from the same or closed related species can be represented as an abelian group, which will be, in turns, the direct sum of abelian cyclic groups of prime-power order. Hence, results lead us to the representation of whole genomes populations of individuals from the same species or close related species (as suggested in Fig.1) by means of direct sum of their group representation into abelian cyclic groups. A general illustration of this modelling would be, for example:

$$S = \ldots \oplus (\mathbb{Z}_{5^3})^{n_1} \oplus \overbrace{(\mathbb{Z}_{2^6})^{m_1}}^{motif} \oplus (\mathbb{Z}_{5^3})^{n_2} \oplus \ldots \oplus \overbrace{(\mathbb{Z}_2^2)^{m_2}}^{domain} \oplus \ldots \oplus \overbrace{(\mathbb{Z}_{5^3})^{n_p}}^{domain} \oplus \overbrace{(\mathbb{Z}_{2^6})^{m_p}}^{motif} \ldots [7]$$

That is, the fundamental theorem of abelian finite groups has an equivalent in genomics.

**Theorem 1**. The genomic architecture from a genome population can be quantitatively represented as an abelian group isomorphic to a direct sum of cyclic groups of prime-power order.

The proof of this theorem is self-evident across the discussion and examples presented here. Basically, the group representations of the genetic code lead to the group representations of local genomic domains in terms of cyclic groups of prime-power order, for example, $(\mathfrak{B}^3,+) \cong (\mathbb{Z}_5^3,+)$ or $(C_e,+) \cong (\mathbb{Z}_{5^3},+)$, till covering the whole genome. As for any finite abelian group, the abelian group representation of genome populations can be expressed in terms a direct sum of abelian cyclic groups of prime-power order. Any new discovering on the annotation of given genome population will only split an abelian group, already defined on some genomic domain/region, into the direct sum of abelian subgroups ■.

# 4 Discussions

Under the assumption that the current forms of life are the result of an evolutionary process started from very simple primordial cells, the current non-coding DNA must be the relict footprint of multiple recombination of ancient DNA domains in all the permissible forms, which in ancient times were rules by an ancient genetic code. In consequence, on this scenario, the group representations of the genetic code are logically extended from relatively small local DNA domains to the whole genome.

Examples shown in Fig. 1 to 4 indicates whatever would be the genomic architecture for given species, the observed variations in the individual populations and in populations from closed related species, it can be quantitatively described as the direct sum of abelian cyclic groups. The discovering/annotation of new genomic features will only lead to the decomposition of previous known abelian cyclic groups representing some genomic subregion into direct sums of subgroups. In such algebraic representation DNA sequence motifs for which only substitution mutations happened are specifically represented by the abelian group $(C_g,+) \cong (\mathbb{Z}_{64},+)$, in protein coding regions, and by any or combination of groups $(B,+) \cong (\mathbb{Z}_2^2,+)$, $(B^2,+) \cong (\mathbb{Z}_2^4,+)$ or $B^3/G_{\text{GGA}} \cong (\mathbb{Z}_2^5,+)$ in non-protein coding regions.

Results indicate that the genome architecture of whole populations can be quantitively studied in the framework of abelian group theory. Two sets of MSA, $S_1$ and $S_2$, could split into different cyclic groups and, however, these sequences can be isomorphic between them because have the same canonical decomposition. Particularly, the genetic code abelian group $(\mathfrak{B}^3,+) \cong (\mathbb{Z}_5^3,+)$ is enough for an algebraic representation of the genome population from the same species or close related species. However, such a decomposition is biologically poor and, as suggested in Figs. 4 to 5, masks relevant biological features from the genome architecture. A further decomposition into the direct sum of abelian groups will only depends on our knowledge on the genome annotation for specified species.

As suggested in Figs. 3 and 4, base D from the extended genetic code (represented as gaps in the MSA) results useful preserving the information on the natural reading frame in the abelian group representation. It is worthy to notices that, for the transcriptional and splicing enzymatic machinery, the information on the reading frame preservation is already in the sequence. Molecular machines perform precise logical operations [39], which in this case result in a sort of molecular *enthymeme* (logical) operation where the conclusion is omitted obeying the principle of cellular economy. In other words, in the algebraic representation of gene and genome populations base D carries real biological information.

From several examples provided here, it is clear that there exists a language for the genome architecture that can be represented in terms of sums of finite abelian groups. The future developments of genome annotation from several species can certainly lead to the discovery of logical rules of a such language determining the possible viable variations in the populations. As suggested in Fig. 5, the identification of quotient groups (at larger scale) can permit the stratification of large genome population into equivalence classes corresponding to individual subpopulations, each one of them carrying particular viable variations of species genome architecture.

As indicated in reference [12], natural genomic rearrangement like DNA recombination and translocation at structural and functional domain can be represented as group automorphisms and endomorphisms. Biologically, such description corresponds to the fact that the new genetic information is recreated, simply, by way of reorganization of the genetic material in the chromosomes of living organisms [5,40]. The analysis and discussion on the application of the endomorphism ring theory to describe the dynamics of genome population is a promising subject for further studies.

Particularly promising is the application of the genomic abelian groups on epigenomic studies, which results when base D stands for the methylated cytosine. As suggested in Fig.5, a precise decomposition of methylation motif into the direct sum of abelian finite group can leads to their classification into unambiguous equivalence classes. This open the doors for the application of based machine-learning bioinformatic approaches to study the methylation changes induced on individual populations by, e.g., environmental changes, aging process and diseases, which is of particular interest in genomic medicine [41].

Results presented here would have considerable positive impact on current molecular evolutionary biology, which heavily relies on evolutionary null hypotheses about the past. As suggested in reference [30], the genomic abelian groups open new horizons for the study of the molecular evolutionary stochastic processes (at genomic scale) and with relevant biomedical applications, founded on a deterministic ground, which only depends on the physicochemical properties of DNA bases and aminoacids. In this case, the only molecular evolutionary hypothesis needed about the past is a fact, the existence of the genetic code.

## 5 Conclusions

Results to date indicate that the genetic code and, ultimately, the physicochemical properties of DNA bases on which the genetic code algebraic structure are defined, has a deterministic effect or at least partially rules on the current genome architectures, in such a way that the abelian group representations of the genetic code are logically extended to the whole genome. In consequence, the fundamental theorem of abelian finite groups can be applied to the whole genome. This result opens new horizons for further genomics studies with the application of the abelian group theory, which currently is well developed and well documented [32,42].

Results suggest that the architecture of current population genomes is quite far from randomness and obeys deterministic rules. Although the random nature of the mutational process, only a small fraction of mutations is fixed in genomic populations. In particular, fixation events are ruled by the genetic code architecture, which as shown by Sanchez (2018), it can be simulated as an optimization process by using genetic algorithms [30]. This points to the study of the dynamics of genome populations as a stochastic deterministic process. Genome stochasticity derives from the stochasticity of mutational process and from the stochasticity of biochemical reactions, which gives rise to a rich population diversity and phenotypic plasticity that help to prevent population extinction. The deterministic part derives from its architecture, which can be represented in terms of a canonical direct sum of homocyclic abelian groups derived from the genetic code, hold for all the individuals from the same population/species.

## References

1. Sanchez R, Barreto J, Morgado E, Grau R. Abelian Finite Group of DNA Genomic Sequences. arXiv Quant Methods. 2005; 1-6. https://arxiv.org/abs/q-bio/0512042. Available: https://www.researchgate.net/publication/2183122_Abelian_Finite_Group_of_DNA_Genomic_Sequences
2. Whalen S, Truty RM, Pollard KS. Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. Nat Genet. 2016;48: 488–496. doi:10.1038/ng.3539
3. Nuebler J, Fudenberg G, Imakaev M, Abdennur N, Mirny LA. Chromatin organization by an interplay of loop extrusion and compartmental segregation. Proc Natl Acad Sci U S A.

2018;115: E6697–E6706. doi:10.1073/pnas.1717730115

4. Rowley MJ, Corces VG. Organizational principles of 3D genome architecture. Nat Rev Genet. Springer US; 2018;19: 789–800. doi:10.1038/s41576-018-0060-8

5. Piazza A, Heyer WD. Homologous Recombination and the Formation of Complex Genomic Rearrangements. Trends Cell Biol. Elsevier Ltd; 2019;29: 135–149. doi:10.1016/j.tcb.2018.10.006

6. Zheng H, Xie W. The role of 3D genome organization in development and cell differentiation. Nat Rev Mol Cell Biol. Springer US; 2019;20: 535–550. doi:10.1038/s41580-019-0132-4

7. Schneider TD. Evolution of biological information. Nucleic Acids Res. 2000;28: 2794–9. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=102656&tool=pmcentrez&rendertype=abstract

8. Yockey HP. Origin of life on earth and Shannon's theory of communication. Comput Chem. 2000;24: 105–123. doi:10.1016/S0097-8485(99)00050-9

9. Sanchez R, Grau R. A genetic code Boolean structure. II. The genetic information system as a Boolean information system. Bull Math Biol. 2005/07/07. 2005;67: 1017–1029. doi:10.1016/j.bulm.2004.12.004

10. Sanchez R, Mackenzie SA. Information thermodynamics of cytosine DNA methylation. Bardoni B, editor. PLoS One. Public Library of Science; 2016;11: e0150427. doi:10.1371/journal.pone.0150427

11. Sánchez R, Morgado E, Grau R. A genetic code Boolean structure. I. The meaning of Boolean deductions. Bull Math Biol. 2005;67: 1–14. doi:10.1016/j.bulm.2004.05.005

12. Sanchez R, Morgado E, Grau R. Gene algebra from a genetic code algebraic structure. J Math Biol. 2005/07/14. 2005;51: 431–457. doi:10.1007/s00285-005-0332-8

13. Sanchez R, Morgado E, Grau R, Sánchez R. A genetic code Boolean structure. I. The meaning of Boolean deductions. Bull Math Biol. 2005/02/05. 2005;67: 1–14. doi:10.1016/j.bulm.2004.05.005

14. Sanchez R, Grau R, Morgado E. A novel Lie algebra of the genetic code over the Galois field of four DNA bases. Math Biosci. 2006;202: 156–174. doi:10.1016/j.mbs.2006.03.017

15. José M V., Zamudio GS, Morgado ER. A unified model of the standard genetic code. R Soc Open Sci. 2017;4: 1–13. doi:10.1098/rsos.160908

16. José M V., Morgado ER, Govezensky T. Genetic Hotels for the Standard Genetic Code: Evolutionary Analysis Based upon Novel Three-Dimensional Algebraic Models. Bull Math Biol. 2011;73: 1443–1476. doi:10.1007/s11538-010-9571-y

17. Sánchez R, Grau R. An algebraic hypothesis about the primeval genetic code architecture. Math Biosci. 2009/07/18. 2009;221: 60–76. doi:S0025-5564(09)00114-X [pii]

10.1016/j.mbs.2009.07.001

18. Orgel LE. Prebiotic chemistry and the origin of the RNA world. Crit Rev Biochem Mol Biol. 2004;39: 99–123. doi:10.1080/10409230490460765

19. Piccirilli JA, Benner SA, Krauch T, Moroney SE, Benner SA. Enzymatic incorporation of a new base pair into DNA and RNA extends the genetic alphabet. Nature. 1990;343: 33–37. doi:10.1038/343033a0

20. Switzer C, Moronev SE, Benner SA. Enzymatic Incorporation of a New Base Pair into DNA and RNA. J Am Chem Soc. 1989;111: 8322–8323. doi:10.1021/ja00203a067

21. Sanchez R, Grau R, Morgado E. A Novel DNA Sequence Vector Space over an extended Genetic Code Galois Field. MATCH Commun Math Comput Chem. 2006;56: 5–20. Available: http://match.pmf.kg.ac.rs/electronic_versions/Match56/n1/match56n1_5-20.pdf

22. Sanchez R. Evolutionary Analysis of DNA-Protein-Coding Regions Based on a Genetic Code Cube Metric. Curr Top Med Chem. 2014;14: 407–417. doi:10.2174/1568026613666131204110022

23. Di Giulio M. LUCA as well as the ancestors of archaea, bacteria and eukaryotes were progenotes: Inference from the distribution and diversity of the reading mechanism of the AUA and AUG codons in the domains of life. BioSystems. Elsevier B.V.; 2020;198: 104239. doi:10.1016/j.biosystems.2020.104239

24. Di Giulio M. Errors of the ancestral translation, LUCA, and nature of its direct descendants. BioSystems. Elsevier B.V.; 2021;206: 104433. doi:10.1016/j.biosystems.2021.104433

25. Smith ZD, Meissner A. DNA methylation: Roles in mammalian development. Nat Rev Genet. Nature Publishing Group; 2013;14: 204–220. doi:10.1038/nrg3354

26. Severin PMD, Zou X, Gaub HE, Schulten K. Cytosine methylation alters DNA mechanical properties. Nucleic Acids Res. 2011;39: 8740–51. doi:10.1093/nar/gkr578

27. Sriraman A, Debnath TK, Xhemalce B, Miller KM. Making it or breaking it: DNA methylation and genome integrity. Essays Biochem. 2020; 687–703. doi:10.1042/ebc20200009

28. Fuchs L. Abelian groups. Publishing House of the Hungarian Academy of Sciences. Publishing House of the Hungarian Academy of Sciences; 1958.

29. Crick FHC. The Origin of the Genetic Code. J Mol Biol. 1968;38: 367–379.

30. Sanchez R. Symmetric Group of the Genetic-Code Cubes. Effect of the Genetic-Code Architecture on the Evolutionary Process. MATCH Commun Math Comput Chem. 2018;79: 527–560. Available: http://match.pmf.kg.ac.rs/electronic_versions/Match79/n3/match79n3_527-560.pdf

31. José M V, Morgado ER, Sánchez R, Govezensky T. The 24 Possible Algebraic Representations of the Standard Genetic Code in Six or in Three Dimensions. Adv Stud Biol.

2012;4: 119–152. Available: http://www.m-hikari.com/asb/asb2012/asb1-4-2012/joseASB1-4-2012-1.pdf

32. Fuchs L. Infinite Abelian Groups, Volume 1. 1st Editio. Academic Press; 1970.

33. Stergachis AB, Haugen E, Shafer A, Fu W, Vernot B, Reynolds A, et al. Exonic transcription factor binding directs codon choice and affects protein evolution. Science (80- ). 2013;342: 1367–72. doi:10.1126/science.1243490

34. Reyna-Llorens I, Burgess SJ, Reeves G, Singh P, Stevenson SR, Williams BP, et al. Ancient duons may underpin spatial patterning of gene expression in C 4 leaves. Proc Natl Acad Sci U S A. 2018;115: 1931–1936. doi:10.1073/pnas.1720576115

35. Yadav VK, Smith KS, Flinders C, Mumenthaler SM, De S. Significance of duon mutations in cancer genomes. Sci Rep. Nature Publishing Group; 2016;6: 1–9. doi:10.1038/srep27437

36. Pujato M, Kieken F, Skiles AA, Tapinos N, Fiser A. Prediction of DNA binding motifs from 3D models of transcription factors; identifying TLX3 regulated genes. Nucleic Acids Res. 2014;42: 13500–13512. doi:10.1093/nar/gku1228

37. Yang M, Derbyshire MK, Yamashita RA, Marchler-Bauer A. NCBI's Conserved Domain Database and Tools for Protein Domain Analysis. Curr Protoc Bioinforma. 2020;69: 1–25. doi:10.1002/cpbi.90

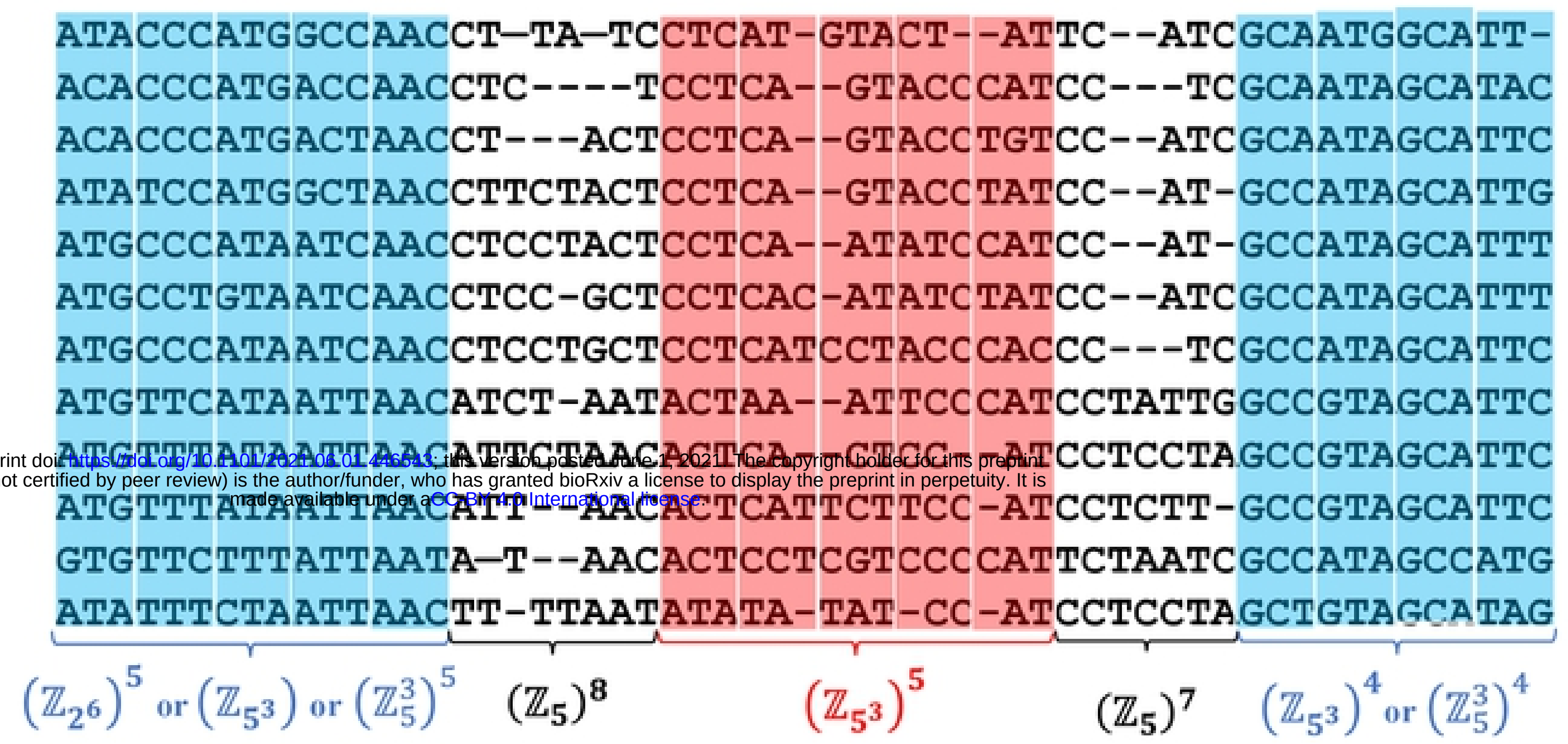
38. Sanchez R, Mackenzie SA. Integrative Network Analysis of Differentially Methylated and Expressed Genes for Biomarker Identification in Leukemia. Sci Rep. 2020;10: 2123. doi:10.1038/s41598-020-58123-2

39. Schneider TD. Theory of molecular machines. II. Energy dissipation from molecular machines. J Theor Biol. 1991;148: 125–137. Available: http://www.ncbi.nlm.nih.gov/pubmed/2016881

40. Yu M, Ren B. The three-dimensional organization of mammalian genomes. Annu Rev Cell Dev Biol. 2017;33: 265–289. doi:10.1146/annurev-cellbio-100616-060531

41. Salameh Y, Bejaoui Y, El Hajj N. DNA Methylation Biomarkers in Aging and Age-Related Diseases. Front Genet. 2020;11: 1–11. doi:10.3389/fgene.2020.00171

42. Fuchs L. Infinite Abelian Groups, Volume 2. Academic Press; 1973.



```
ATACCCATGGCCAACCT—TA—TCCTCAT-GTACT--ATTC--ATCGCAATGGCATT-
ACACCCATGACCAACCTC----TCCTCA--GTACCCATCC---TCGCAATAGCATAC
ACACCCATGACTAACCT---ACTCCTCA--GTACCTGTCC--ATCGCAATAGCATTC
ATATCCATGGCTAACCTTCTACTCCTCA--GTACCTATCC--AT-GCCATAGCATTG
ATGCCCATAATCAACCTCCTACTCCTCA--ATATCCATCC--AT-GCCATAGCATTT
ATGCCTGTAATCAACCTCC-GCTCCTCAC-ATATCTATCC--ATCGCCATAGCATTT
ATGCCCATAATCAACCTCCTGCTCCTCATCCTACCCACCC---TCGCCATAGCATTC
ATGTTCATAATTAACATCT-AATACTAA--ATTCCCATCCTATTGGCCGTAGCATTC
ATGTTTATAATTAACATTCTAACACTCA--GTCC--ATCCTCCTAGCCGTAGCATTC
ATGTTTATAATTAACATT--AACACTCATTCTTCC-ATCCTCTT-GCCGTAGCATTC
GTGTTCTTTATTAATA—T--AACACTCCTCGTCCCCATTCTAATCGCCATAGCCATG
ATATTTCTAATTAACTT-TTAATATATA-TAT-CC-ATCCTCCTAGCTGTAGCATAG
```

$(\mathbb{Z}_{2^6})^5$ or $(\mathbb{Z}_{5^3})$ or $(\mathbb{Z}_5^3)^5$
$(\mathbb{Z}_5)^8$
$(\mathbb{Z}_{5^3})^5$
$(\mathbb{Z}_5)^7$
$(\mathbb{Z}_{5^3})^4$ or $(\mathbb{Z}_5^3)^4$

Figure 1

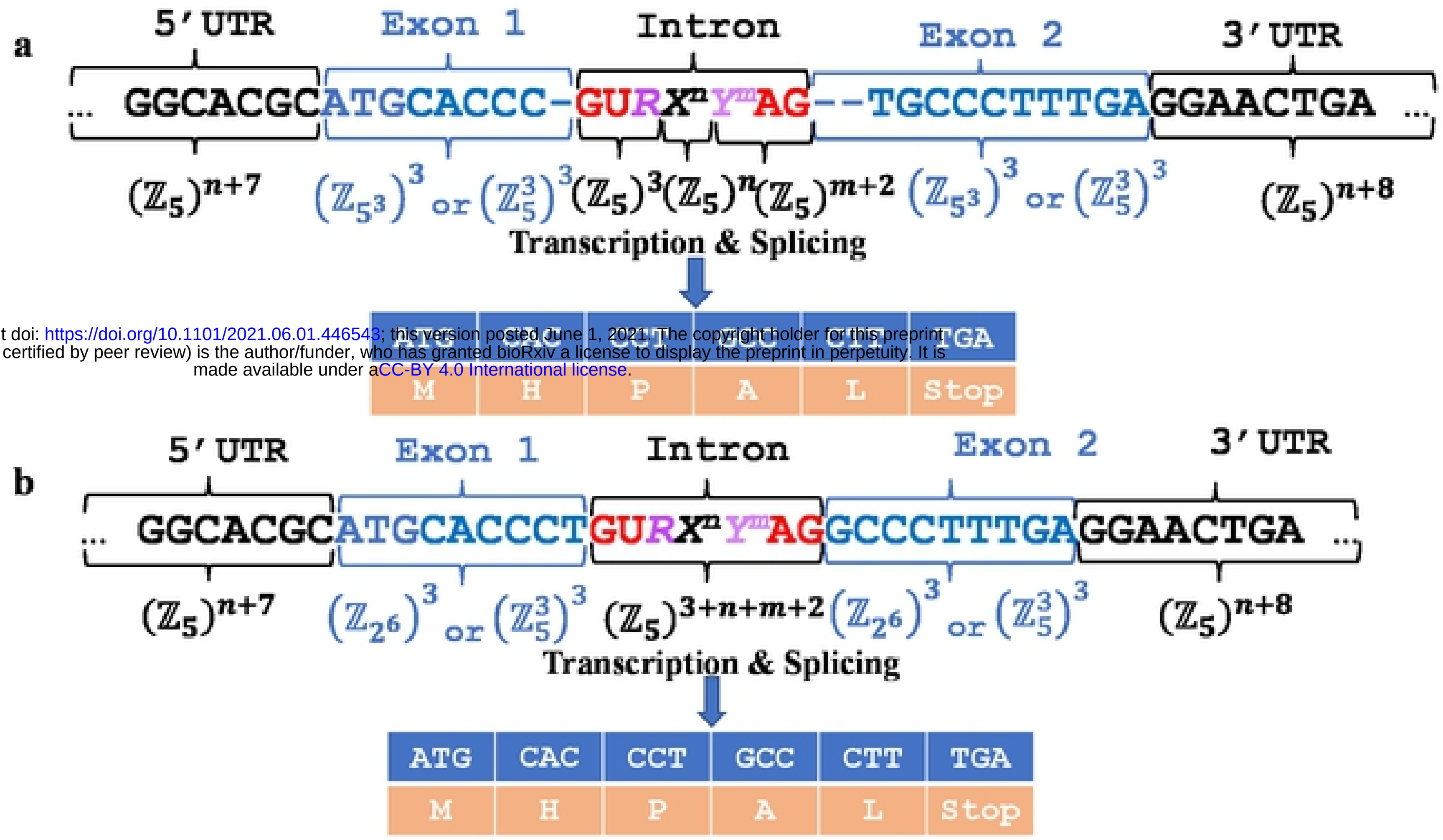
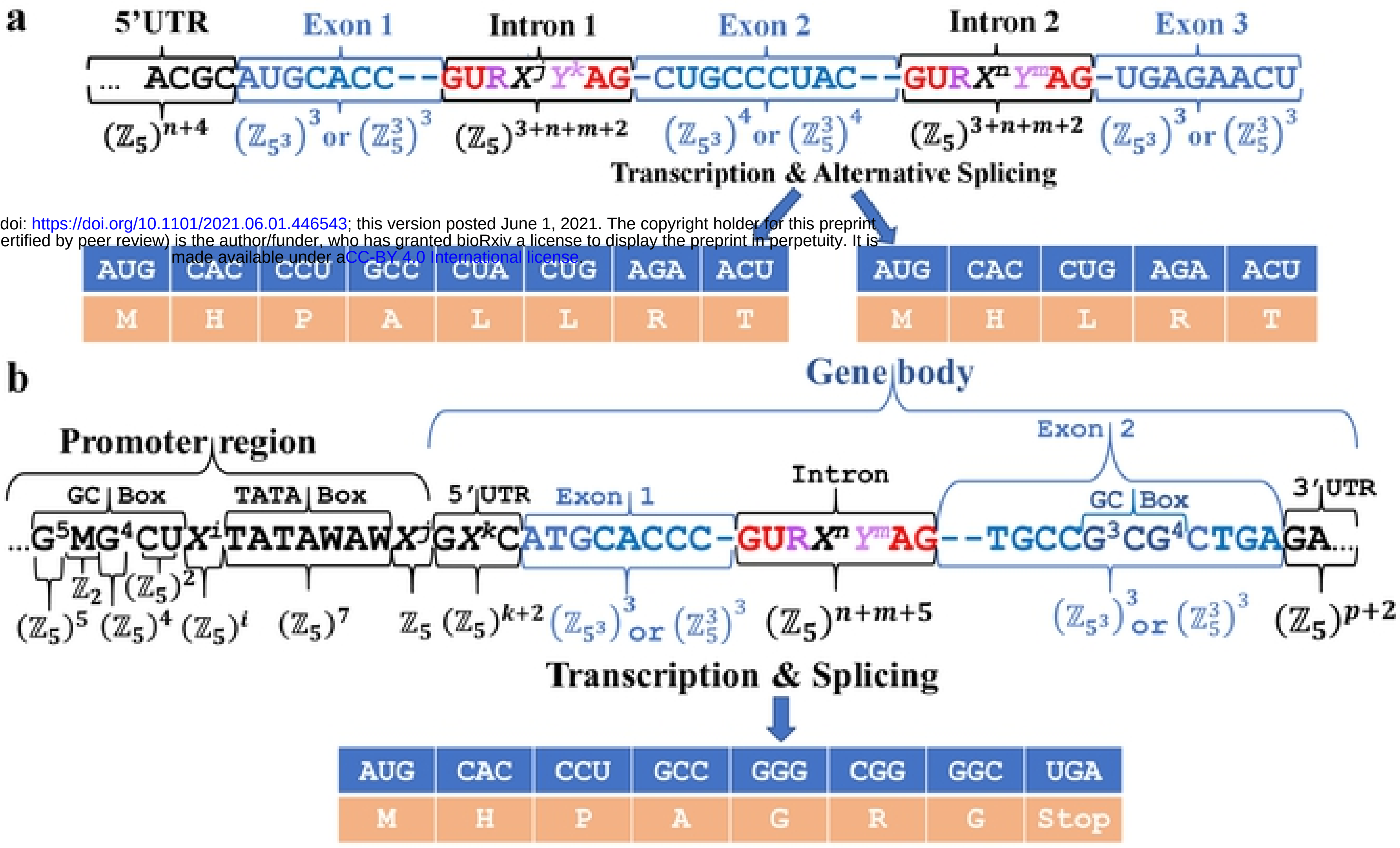

a
5′ UTR
Exon 1
Intron
Exon 2
3′ UTR
... GGCACGCATGCACCC-GURX$^n$Y$^m$AG--TGCCCTTTGAGGAACTGA ...
$(\mathbb{Z}_5)^{n+7}$
$(\mathbb{Z}_{5^3})^3$ or $(\mathbb{Z}_5^3)^3$
$(\mathbb{Z}_5)^3$
$(\mathbb{Z}_5)^n$
$(\mathbb{Z}_5)^{m+2}$
$(\mathbb{Z}_{5^3})^3$ or $(\mathbb{Z}_5^3)^3$
$(\mathbb{Z}_5)^{n+8}$
Transcription & Splicing
TGA
M
H
P
A
L
Stop
b
5′ UTR
Exon 1
Intron
Exon 2
3′ UTR
... GGCACGCATGCACCCTGURX$^n$Y$^m$AGGCCCTTTGAGGAACTGA ...
$(\mathbb{Z}_5)^{n+7}$
$(\mathbb{Z}_{2^6})^3$ or $(\mathbb{Z}_5^3)^3$
$(\mathbb{Z}_5)^{3+n+m+2}$
$(\mathbb{Z}_{2^6})^3$ or $(\mathbb{Z}_5^3)^3$
$(\mathbb{Z}_5)^{n+8}$
Transcription & Splicing
ATG
CAC
CCT
GCC
CTT
TGA
M
H
P
A
L
Stop

Figure 3


a
5'UTR
Exon 1
Intron 1
Exon 2
Intron 2
Exon 3
... ACGCAUGCACC--GURX$^j$ Y$^k$AG-CUGCCCUAC--GURX$^n$ Y$^m$AG-UGAGAACU
$(\mathbb{Z}_5)^{n+4}$
$(\mathbb{Z}_{5^3})^3$ or $(\mathbb{Z}_5^3)^3$
$(\mathbb{Z}_5)^{3+n+m+2}$
$(\mathbb{Z}_{5^3})^4$ or $(\mathbb{Z}_5^3)^4$
$(\mathbb{Z}_5)^{3+n+m+2}$
$(\mathbb{Z}_{5^3})^3$ or $(\mathbb{Z}_5^3)^3$
Transcription & Alternative Splicing
b
Gene body
Exon 2
Promoter region
GC Box
TATA Box
5'UTR
Exon 1
Intron
GC Box
3'UTR
...G$^5$MG$^4$CUX$^i$TATAWAWX$^j$GX$^k$CATGCACCC-GURX$^n$ Y$^m$AG--TGCCG$^3$CG$^4$CTGAGA...
$\mathbb{Z}_2$
$(\mathbb{Z}_5)^2$
$(\mathbb{Z}_5)^5$
$(\mathbb{Z}_5)^4$
$(\mathbb{Z}_5)^i$
$(\mathbb{Z}_5)^7$
$\mathbb{Z}_5$
$(\mathbb{Z}_5)^{k+2}$
$(\mathbb{Z}_{5^3})^3$ or $(\mathbb{Z}_5^3)^3$
$(\mathbb{Z}_5)^{n+m+5}$
$(\mathbb{Z}_{5^3})^3$ or $(\mathbb{Z}_5^3)^3$
$(\mathbb{Z}_5)^{p+2}$
Transcription & Splicing

| AUG | CAC | CCU | GCC | CUA | CUG | AGA | ACU |
|---|---|---|---|---|---|---|---|
| M | H | P | A | L | L | R | T |

| AUG | CAC | CUG | AGA | ACU |
|---|---|---|---|---|
| M | H | L | R | T |

| AUG | CAC | CCU | GCC | GGG | CGG | GGC | UGA |
|---|---|---|---|---|---|---|---|
| M | H | P | A | G | R | G | Stop |

Figure 4

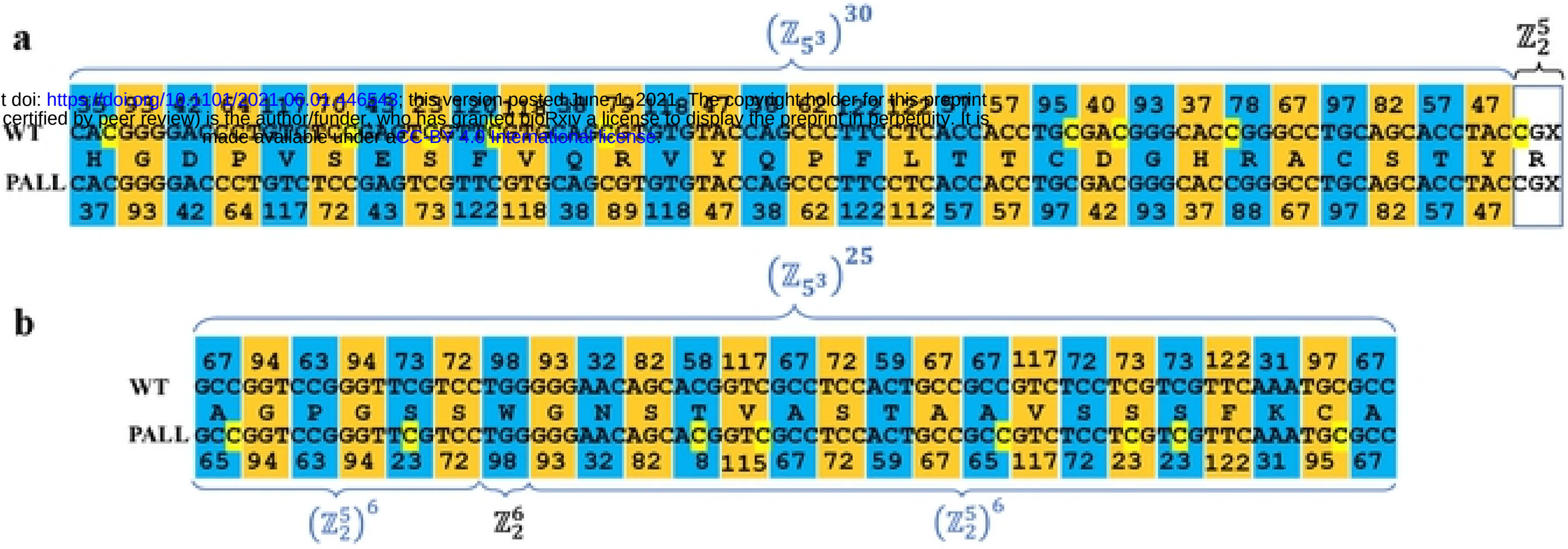

a
$(\mathbb{Z}_{5^3})^{30}$
$\mathbb{Z}_2^5$
WT
57 95 40 93 37 78 67 97 82 57 47
ACCTGCGACGGGCACCGGGCCTGCAGCACCTACCGX
H G D P V S E S F V Q R V Y Q P F L T T C D G H R A C S T Y R
PALL
CACGGGGACCCTGTCTCCGAGTCGTTCGTGCAGCGTGTGTACCAGCCCTTCCTCACCACCTGCGACGGGCACCGGGCCTGCAGCACCTACCGX
37 93 42 64 117 72 43 73 122 118 38 89 118 47 38 62 122 112 57 57 97 42 93 37 88 67 97 82 57 47
b
$(\mathbb{Z}_{5^3})^{25}$
67 94 63 94 73 72 98 93 32 82 58 117 67 72 59 67 67 117 72 73 73 122 31 97 67
WT
GCCGGTCCGGGTTCGTCCTGGGGGAACAGCACGGTCGCCTCCACTGCCGCCGTCTCCTCGTCGTTCAAATGCGCC
A G P G S S W G N S T V A S T A A V S S S F K C A
PALL
GCCGGTCCGGGTTCGTCCTGGGGGAACAGCACGGTCGCCTCCACTGCCGCCGTCTCCTCGTCGTTCAAATGCGCC
65 94 63 94 23 72 98 93 32 82 8 115 67 72 59 67 65 117 72 23 23 122 31 95 67
$(\mathbb{Z}_2^5)^6$
$\mathbb{Z}_2^6$
$(\mathbb{Z}_2^5)^6$

Figure 5

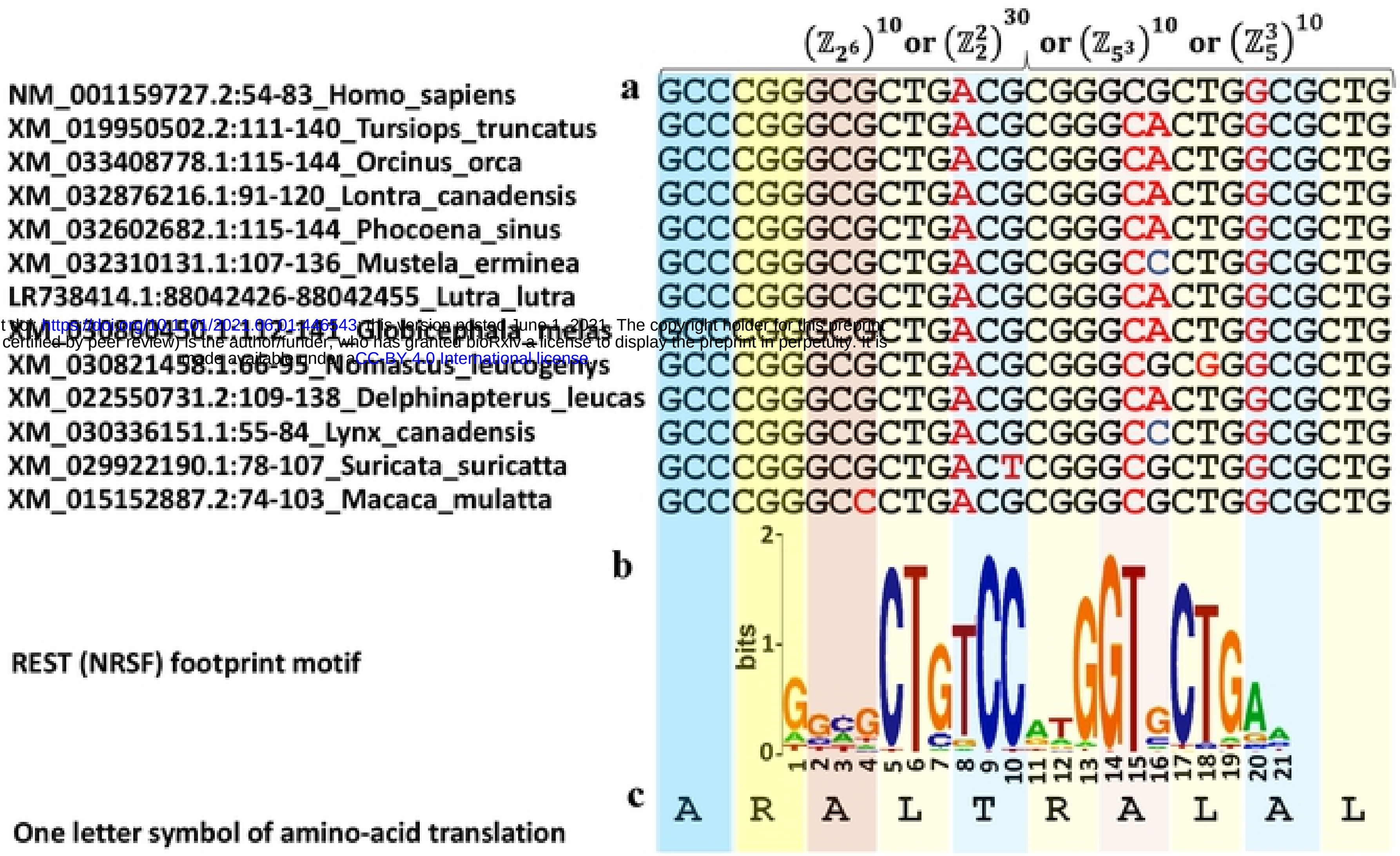

$(\mathbb{Z}_{2^6})^{10}$ or $(\mathbb{Z}_2^2)^{30}$ or $(\mathbb{Z}_{5^3})^{10}$ or $(\mathbb{Z}_5^3)^{10}$
a
NM_001159727.2:54-83_Homo_sapiens GCCCGGGCGCTGACGCGGGCGCTGGCGCTG
XM_019950502.2:111-140_Tursiops_truncatus GCCCGGGCGCTGACGCGGGCACTGGCGCTG
XM_033408778.1:115-144_Orcinus_orca GCCCGGGCGCTGACGCGGGCACTGGCGCTG
XM_032876216.1:91-120_Lontra_canadensis GCCCGGGCGCTGACGCGGGCACTGGCGCTG
XM_032602682.1:115-144_Phocoena_sinus GCCCGGGCGCTGACGCGGGCACTGGCGCTG
XM_032310131.1:107-136_Mustela_erminea GCCCGGGCGCTGACGCGGGCCCTGGCGCTG
LR738414.1:88042426-88042455_Lutra_lutra GCCCGGGCGCTGACGCGGGCACTGGCGCTG
XM_030860450.1:112-141_Globicephala_melas GCCCGGGCGCTGACGCGGGCACTGGCGCTG
XM_030821458.1:66-95_Nomascus_leucogenys GCCCGGGCGCTGACGCGGGCGCGGGCGCTG
XM_022550731.2:109-138_Delphinapterus_leucas GCCCGGGCGCTGACGCGGGCACTGGCGCTG
XM_030336151.1:55-84_Lynx_canadensis GCCCGGGCGCTGACGCGGGCCCTGGCGCTG
XM_029922190.1:78-107_Suricata_suricatta GCCCGGGCGCTGACTCGGGCGCTGGCGCTG
XM_015152887.2:74-103_Macaca_mulatta GCCCGGGCCCTGACGCGGGCGCTGGCGCTG
b
REST (NRSF) footprint motif
bits
2
1
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
c
One letter symbol of amino-acid translation
A R A L T R A L A L

Figure 2