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 DNA sequence domains suggesting a well-structured genomic architecture, where, for each species, genome populations are integrated by individual mutational variants. Herein, we show that, consistent with the fundamental theorem of Abelian finite groups, 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 of prime-power order defined on the genetic code and on the set of DNA bases, where populations can be stratified into subpopulations with the same canonical decomposition into *p*-groups. Through concrete examples we show that the architectures of current annotated genomic regions including (but not limited to) transcription factors binding-motif, promoter regulatory boxes, exon and intron arrangement associated to gene splicing are subjects for feasible modeling as decomposable Abelian *p*-groups. Moreover, we show that the epigenomic variations induced by diseases or environmental changes also can be represented as an Abelian group decomposable into homocyclic Abelian *p*-groups. The nexus between the direct sum of homocycle Abelian *p*-groups and the endomorphism ring paved the ways to unveil unsuspected stochastic-deterministic logical propositions ruling the ensemble of genomic regions. Our study aims to set the basis for concrete applications of the theory in computational biology and bioinformatics. Consistently with this goal, a computational tool designed for the analysis of fixed mutational events in gene/genome populations represented as endomorphisms and automorphisms is provided. Results suggest that complex local architectures and evolutionary features no evident through the direct experimentation can be unveiled through the analysis of the endomorphism ring and the subsequent application of machine learning approaches for the identification of stochastic-deterministic logical rules (reflecting the evolutionary pressure on the region) constraining the set of possible mutational events (represented as homomorphisms) and the evolutionary paths.

Keywords: Genomics, Genetic code, Abelian groups, genome algebra, automorphism, mutational event

1 Introduction

The analysis of the *genome architecture* is one of biggest challenges for the current and future genomics. Herein, with the term *genome architecture* we are adopting the definition given by Koonin [1]: *Genome architecture can be defined as the totality of non-random arrangements of functional elements (genes, regulatory regions, etc.) in the genome.*

Current bioinformatic tools make possible faster genome annotation process (identification of locations for genes, regulatory regions, intron-exon boundaries, repeats, etc.) 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].

We propose the study of genome architecture in the context of population genomics, where all the variability constrained by the evolutionary pressure is expressed. Although the random nature of the mutational process, only a small fraction of mutations is fixed in genomic populations. In particular, fixation events, ultimately guided by random genetic drift and positive selection are constrained by the genetic code, which permits a probabilistic estimation of the evolutionary mutational cost by simulating the evolutionary process as an optimization process with genetic algorithms [11].

1.1 The genetic code

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 is the cornerstone of live on earth, the fundamental communication code from the genetic information system [8,9]. The code-words from the genetic code are given in the alphabet of four DNA bases $\mathfrak{B} = \{\text{A}, \text{C}, \text{G}, \text{T}\}$ and integrates a set of 64 DNA base-triplets $\{XYZ\}$ also named *codons*, where $X, Y, Z \in \mathfrak{B}$. Each codon encodes the information for one aminoacids and each aminoacid is encoded by one or more codons. Hence, at biomolecular level, the genetic code constitute a set of biochemical rules (mathematically expressed as an injective mapping: *codon* → *aminoacid*) used by living cells to translate information encoded within genetic material into proteins, which sets the basis for our understanding of the mathematical logic inherent to the genetic information system [9,12].

The subjacent idea to impose a group structure on the set of codons resides on that the genetic code is the code of a communication system, the genetic information system [8,13,14]. As suggested by Andrews and Boss [15]: "*In codes used for electrical transmission of engineering signals, **group structure is imposed to increase efficiency and reduce error**. Similarly, the group characteristics of codon redundancy could serve to transmit additional information superimposed on the messages directing amino acid order in protein synthesis"*. As in the current human communication systems [16], to impose a group structure (on biophysical basis) on the set of codons facilitate a better understanding and evaluation of the error performance and efficiency of the genetic message carried in the chromosomes across generations [15].

1.2 The genetic code algebraic structures

The basis of the current study are algebraic structures (specifically groups structures) defined on the set of bases and on the codon sets. We assume that readers are familiar with algebraic structures like group, ring, and the classical mapping defined on them, homomorphisms, automorphism, and translations. For readers not familiar with this subject, a brief basic introduction to these definitions is given in the Appendix.

The meaning of group operations. Group operations are defined on the sets of DNA bases and codons, are associated to physicochemical or/and biophysical relationships between DNA bases and between codons and aminoacids. In other words, a proper definition of a group operation on the set of bases or on the set of codons will encode the physicochemical or/and biophysical relationships between the set's elements. Thus, by group operations defined on the set of bases or on the set of codons, we understand an *encoding* applied to represent specified physicochemical or/and biophysical relationships as group operations between the elements of the set. Then, we shall say that such an encoding permits the *representation* of DNA bases, codons, genes, and genomic sequences as elements from algebraic structures.

Obviously, depending on which physicochemical or biophysical relationship is under scrutiny, different encodings of the group operations can be defined on the sets of bases and codons, as shown in reference [17]. The meaning of the group operations has been subjects of the references where the corresponding groups have been reported [11,17–20]. For example, in the DNA double helix, nucleotide bases are paired following specific physicochemical relationships: 1) *the chemical type sets the main rule for a paring*: *a purine base is paired with a pyrimidine*, 2) *paired bases must have the same hydrogen-bonding capability*. These physicochemical relationships rule the DNA base pairing: G:::C (*three hydrogen bonds*) and A::T (*two hydrogen bonds)*. In this scenario, the sum operation is defined in [20], over the ordered set of bases $\mathfrak{B} = \{D, \mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$, in such a way that the DNA complementary bases are also complementary algebraic elements.

Pioneering works on the genetic code algebraic structure. Pioneering works were made in the 70s [15,21–23], just few years after Nirenberg won the Nobel Prize in Physiology or Medicine (in 1968) for his seminal work on the genetic code. Andrews and Boss proposed the cyclic groups of DNA bases, which is isomorphic to the Abelian group defined on the set of integers modulo 4, $\mathbb{Z}_4(\mathbb{Z}/4\mathbb{Z}$) [15]. Their approach also considered the base representation with cyclic group of complex numbers. Further studies were focused on operational groups applied to transform bases and base-doublet into each other. Dankworth and Neubert (1975) proposed the Klein-4-group structure (*K*) of doublet-exchange operators and applied the direct product *K*×*K* to study the symmetries of genetic-code doublets [22]. The four dimensional hypercube structure of the genetic-code doublets (*K*×*K* group) was later studied by Bergman and Jungck (1979) [23].

Efforts with the application of *group representation theory* to study the origin and evolution of the genetic code were made by Honors and Hornos [24,25], and extended to Lie superalgebras by Forger and Sachse [26]. However, these efforts on the application of group representation theory are heavily relying on physical interpretations disconnected from concrete molecular biology context, which made hard a further application on concrete molecular biology or computational biology studies, and on bioinformatic applications. Here, it is important to recall that the *representation* of DNA bases, codons, genes, and genomic sequences as elements from algebraic structures must not be confused with the term *group representation* typically used in algebra referring to the theory of representations of algebraic structures or, particularly, the *group representation theory*. Nevertheless, once a group structure has been defined, for example, in the set of codons, a further application of the group representation theory can be developed.

In the current study, we aim to show that all possible genomic regions and, consequently, whole chromosomes 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. Concepts and basic applications are introduced step by step, sometimes with self-evident statements for a reader familiar with molecular biology. However, it will be shown that the algebraic modeling is addressed to unveil more complex relationships between molecular evolutionary process and the genomic architecture than those eyes-visible relationships. This goal will be evidenced on section 3.2. Our algebraic model approach is intended to set the theoretical basis for further studies addressed to unveil and to understand the rules on how genomes are built. Concrete examples and an implementation in a R package are provided to pave the way for future computational and bioinformatic applications. A graphical summary of the modeling of DNA genomic regions proposed here is shown in Fig 1.

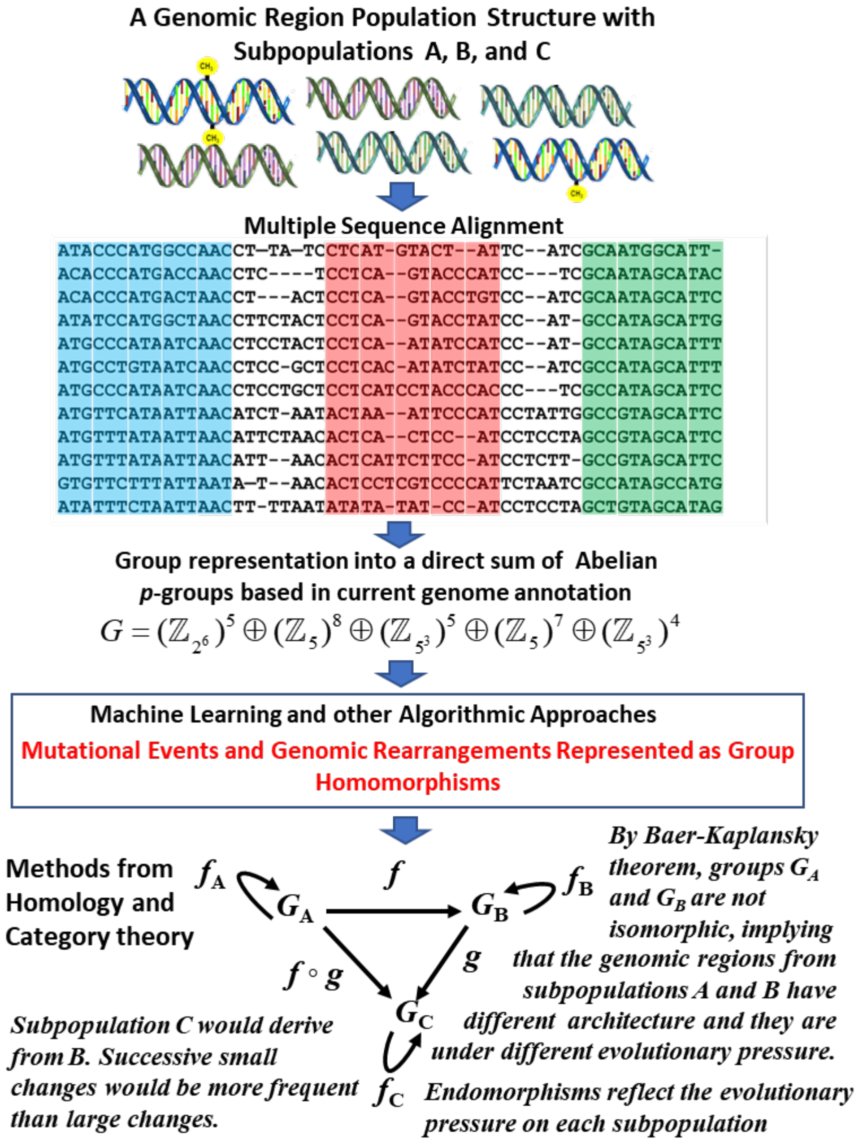

A Genomic Region Population Structure with Subpopulations A, B, and C
CH3
CH3
CH3
Multiple Sequence Alignment
ATACCCATGGCCAACCT—TA—TCCTCAT-GTACT--ATTC--ATCGCAATGGCATT-
ACACCCATGACCAACCTC----TCCTCA--GTACCCATCC---TCGCAATAGCATAC
ACACCCATGACTAACCT---ACTCCTCA--GTACCTGTCC--ATCGCAATAGCATTC
ATATCCATGGCTAACCTTCTACTCCTCA--GTACCTATCC--AT-GCCATAGCATTG
ATGCCCATAATCAACCTCCTACTCCTCA--ATATCCATCC--AT-GCCATAGCATTT
ATGCCTGTAATCAACCTCC-GCTCCTCAC-ATATCTATCC--ATCGCCATAGCATTT
ATGCCCATAATCAACCTCCTGCTCCTCATCCTACCCACCC---TCGCCATAGCATTC
ATGTTCATAATTAACATCT-AATACTAA--ATTCCCATCCTATTGGCCGTAGCATTC
ATGTTTATAATTAACATTCTAACACTCA--CTCC--ATCCTCCTAGCCGTAGCATTC
ATGTTTATAATTAACATT--AACACTCATTCTTCC-ATCCTCTT-GCCGTAGCATTC
GTGTTCTTTATTAATA—T--AACACTCCTCGTCCCCATTCTAATCGCCATAGCCATG
ATATTTCTAATTAACTT-TTAATATATA-TAT-CC-ATCCTCCTAGCTGTAGCATAG
Group representation into a direct sum of Abelian *p*-groups based in current genome annotation
$G = (\mathbb{Z}_{2^6})^5 \oplus (\mathbb{Z}_5)^8 \oplus (\mathbb{Z}_{5^3})^5 \oplus (\mathbb{Z}_5)^7 \oplus (\mathbb{Z}_{5^3})^4$
Machine Learning and other Algorithmic Approaches
Mutational Events and Genomic Rearrangements Represented as Group Homomorphisms
Methods from Homology and Category theory
f_A
G_A
f
G_B
f_B
$f \circ g$
g
G_C
f_C
By Baer-Kaplansky theorem, groups G_A and G_B are not isomorphic, implying that the genomic regions from subpopulations A and B have different architecture and they are under different evolutionary pressure.
Subpopulation C would derive from B. Successive small changes would be more frequent than large changes.
Endomorphisms reflect the evolutionary pressure on each subpopulation

Fig 1. Graphical of the summary showing the bioinformatic and analytical steps followed in the algebraic modeling proposed in current work.

2 Materials and Methods

2.1 Preceding models applied in the current work

Of particular interest are the Abelian *p*-groups defined on the set of DNA bases $\mathfrak{B} = \{A, C, G, T\}$ and on the set of 64 codons $C_g = \{XYZ \mid X, Y, Z \in \mathfrak{B}\}$, which are applied to modeling the physicochemical relationships between DNA bases in the codons [11,18]. Herein, for application purposes in computational biology and bioinformatics addressed to the study of the genome architecture, we focused our study on Abelian *p*-groups defined on $\mathfrak{B}$ and on C_g isomorphic to the groups $\mathbb{Z}_{p_i^{\alpha_i}}$, $p_i^{\alpha_i} \in \{2^2, 2^6\}$, and on $\mathfrak{B}_+ = \{\mathrm{A,C,G,T,}D\}$ and $C_{g+} = \{XYZ \mid X,Y,Z \in \mathfrak{B}_+\}$, $p_i^{\alpha_i} \in \{5, 5^3\}$, as presented in references [11,17–20].

Setting different physicochemical restrictions on the definition of groups operations leads to the 24 possible algebraic representations of the genetic code [17]. In particular, the Abelian *p*-group representations on the set $C_G = \mathfrak{B}\times\mathfrak{B}\times\mathfrak{B}$ and $C_{G+} = \mathfrak{B}_+\times\mathfrak{B}_+\times\mathfrak{B}_+$ ($\mathfrak{B}_+ = \{\mathrm{A,C,G,T,}D\}$, where *D* stands for an alternative base, see below) are isomorphic to Abelian groups defined on $\mathbb{Z}^3_{2^2}$ and $\mathbb{Z}^3_5$, respectively. These group structures lead to 24 (isomorphic) geometrical representations of the genetic code as cubes inserted in three-dimensional space [11,17,19,20] (Fig 2 and SI Figs 1 and 3).

As shown in reference [11], a group structure isomorphic to the symmetric group of degree four S_4 (preserving the group operations previously defined on the codon set) can be defined in set the 24 genetic-code algebraic representations or in the set 24 cubes. Since the definition of a sum operation over the base set is equivalent to define an order on it, *cubes are named according to the base order on them.* For example, the cube shown in Fig 2 is denoted as ACGT, which correspond to the group operation defined on the ordered set $\mathfrak{B} = \{\mathrm{A,C,G,T,D}\}$ (the 'dual' cube TGCA is shown in SI Fig 2 [11]). Simulation of the evolutionary mutational process with the application of genetic algorithms indicates that fixed mutational events found in different protein populations are *very restrictive* in the sense that the optimal evolutionary codon distances are reached for specific models of genetic-code cube or for specific combination of genetic-code cube models [11]. In the present work, it will be shown that codon mutational events represented in terms of automorphisms can be also restrictive for specific genetic-code cube models (section 3.1).

All the Abelian *p*-group included in the current work are oriented to the study of the mutational process [11,17–20]. That is, since we are interested in those structures that permit the analysis and quantitative description of the mutational process in organismal populations, where mutational event can be represented by means of endomorphisms, automorphisms, and translations on the defined group, we do not include algebraic structures designed to study the origin and evolution of the genetic code [11,18]. The genetic code is taken as currently is, without over-impose any evolutionary hypothesis on it.

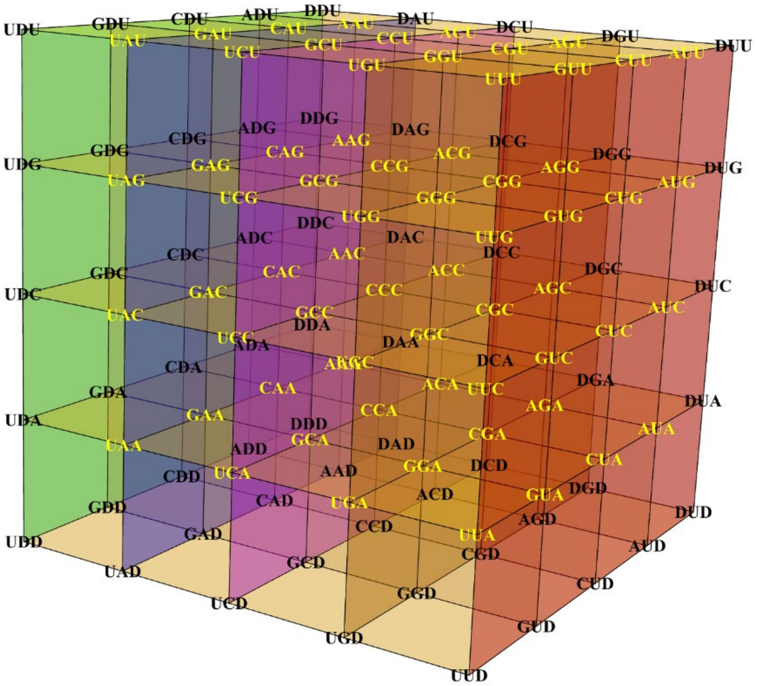


Fig 2. Geometrical representation of the genetic code as a cube inserted in three-dimensional space. The 2-group and 5-group representation defined on the sets $C_G = \mathfrak{B}\times\mathfrak{B}\times\mathfrak{B}$ and $C_{G+} = \mathfrak{B}_+\times\mathfrak{B}_+\times\mathfrak{B}_+$ isomorphic to the groups defined on $\mathbb{Z}_{2^2}^3$ and $\mathbb{Z}_5^3$, respectively, lead to the geometrical representations of the genetic code as a cube inserted in three-dimensional space. The cube corresponding to the base-triplets with coordinates on $\mathbb{Z}_{2^2}^3$ (yellow codons) is inserted in the cube with codon coordinates on $\mathbb{Z}_5^3$. The extended base-triplets including the alternative base D (in black) are located on the cartesian coordinate planes. Codons encoding for amino acids with similar physicochemical properties are located on the same vertical plane (for more details on the cube description see also SI Fig 1 and reference [11,17,19,20]).

A general model also consider Abelian 5-groups that includes a dummy variable (denoted by letter D), which extends the DNA alphabet to five letters. The usefulness of including a fifth base in the evolutionary analysis was shown in reference [20], where two evolutionary models, an algebraic and a stationary Markov (process) models, were applied to phylogenetic analysis reaching (both models) greater discriminatory power than the (now) classical Tamura-Neil evolutionary (Markov) model based on four DNA alphabet [27]. Depending on the concrete application, letter "D" will take a different value. The possible values in the context of the present modeling are: 1) the gap symbol

"-", which stands for insertion deletion/mutations in the multiple sequence alignment (MSA) of DNA sequences, 2) alternative wobble base pairing (e.g., bases such as: inosine (in eukaryotes), agmatine (in archaea), and lysidine (in bacteria) [17,21,22]), and 3) 5-methylcytosine (C^m) and N-6-methyladenine (A^m) when intended for epigenetic studies.

A concrete application of the extended genetic-code cubes over the Galois field *GF*(5) to the simulation of the mutational process proposed in reference [11] would be particularly relevant to predict immunoescape epitope variants originated in populations of pathogenic microorganisms and viruses. In addition, examples provided (here) on the application of the algebraic model to DNA methylation (on 5-methylcytosine and on N-6-methyladenine) suggest its importance for epigenetic studies. The analysis of the fixed mutational events on genes populations revealed that the mutational process can be described by automorphisms on different cubes or sets of cubes [11]. The best genetic-code cubes describing the mutational process on a given gene population are selected with the application of an optimization algorithm (evolutionary (genetic) algorithms) using multiple sequence alignment as raw data [11].

It is worthy to notice that, for all mentioned Abelian *p*-groups, the calculus can be accomplished as symbolic computation on the set of DNA bases or on the set of codons (see e.g., [18]). However, for practical purposes, we take advantage of the group isomorphisms. That is, after define group structures on the sets of bases and codons, for the sake of straightforward computation it is convenient to take advantage of the group isomorphisms with the Abelian *p*-groups like: $\mathbb{Z}_{2^2}$, $\mathbb{Z}_{2^2}^3$, $\mathbb{Z}_{2^6}$, $\mathbb{Z}_5$, $\mathbb{Z}_{5^3}$ and $\mathbb{Z}_5^3$, which will be used in our study instead of the original groups defined on the sets of bases and codons (base-triplets). An introductory summary on the mentioned algebraic structure defined on the set of codons is provided as supporting information in S1.

In the context of genetic-code algebraic structures, by the term "*representation*" of DNA bases, codons, genes, and genomic sequences as elements from algebraic structures, we understand the symbolic representation of the mentioned biomolecules and the physicochemical relationships between them by means of group operations defined on the given set of biomolecules.

2.2 Aligned DNA sequences and data sets

All the DNA sequence alignments and data sets used in this work are available within the R package *GenomAutomorphism* (version 1.0.0) [28]. In addition, the pairwise sequence alignments of SARS coronaviruses used the analyses shown in Fig 8**a** and **b** are also available at GitHub in: https://github.com/genomaths/seqalignments/tree/master/COVID-19. The multiple sequence alignment (MSA) of primate somatic cytochrome c and data description are available on GitHub at: https://github.com/genomaths/seqalignments/tree/master/CYCS. This MSA includes DNA protein-coding sequences from: human, gorilla, silvery gibbon, white cheeked gibbon, Francois langur, olive baboon, golden monkey, rhesus monkeys, gelada baboon, and orangutan. The MSA of primate BRCA1 (transcript variant 4) DNA repair gene used to compute the automorphism shown Fig 8**d** is available on GitHub at https://github.com/genomaths/seqalignments/tree/master/BRCA1. The MSA, coordinates and R script to create the sequence-logo from Fig 4 are given in the Supporting Information.

2.3 Software applied for the mathematical and statistical analyses

Results shown in Fig 8 and Fig 9 were obtained applying the *GenomAutomorphism* R package [28] (version 1.0.0), which is available at Bioconductor (the open source software for Bioinformatics, version: 3.16) and, also, in GitHub at: https://github.com/genomaths/GenomAutomorphism. The whole R script pipeline applied in the estimation of automorphisms (Fig 8) and decision tree (Fig 9) are available as tutorials (vignettes) at the *Geno Automorphism* website: https://github.com/genomaths/GenomAutomorphismm.

The estimation of the best fitted probability distribution shown in Fig 8**f** was accomplished with R package *usefr* available at GitHub: https://github.com/genomaths/usefr, and the goodness-of-fit tests are reported in the mentioned tutorials.

The genetic-code cube shown in Fig 2 was obtained from the Wolfram Mathematica Notebook: *Introduction to $\mathbb{Z}_5$-Genetic-Code vector space*, free available at https://github.com/genomaths/GenomeAlgebra_SymmetricGroup.

2.4 Theoretical Model

According to the fundamental theorem of Abelian finite groups (FTAG) [29,30], any finite Abelian group can be decomposed into a direct sum of homocyclic *p*-groups [29], i.e., a group in which the order of every element is a power of a primer number *p*. Herein, it will be showed that, in a general scenario, genomic regions and, consequently, whole genome populations from any species or close related species, can be algebraically represented as a direct sum of Abelian homocyclic groups or more specifically Abelian *p*-groups of *prime-power order*. The multiple sequence alignments (MSA) of a given genomic region of *N* base-pair (bp) length can be represented as the direct sum:

$$G = \left(\mathbb{Z}_{p_1^{\alpha_1}}\right)^{n_1} \oplus \left(\mathbb{Z}_{p_2^{\alpha_2}}\right)^{n_2} \oplus \cdots \oplus \left(\mathbb{Z}_{p_k^{\alpha_k}}\right)^{n_k} \qquad (1)$$

Where $p_i^{\alpha_i} \in \{2, 5, 2^6, 5^3\}$, n_i stands for the number of cyclic groups $\mathbb{Z}_{p_i^{\alpha_i}}$ integrating the homocyclic group $\left(\mathbb{Z}_{p_i^{\alpha_i}}\right)^{n_i} = \mathbb{Z}_{p_i^{\alpha_i}} \overset{n_i \ times}{\oplus \ldots \oplus} \mathbb{Z}_{p_i^{\alpha_i}}$. Here, we assume the usual definition of direct sum of groups [30]. For $p_j^{\alpha_j} \in \{2^6, 5^3\}$ the cyclic group $\mathbb{Z}_{p_j^{\alpha_j}}$ will cover three bases, otherwise only one base (see examples below). Considering such groups (not necessarily in the order given in Eq. 1) we have: $N = n_1 + \ldots + n_j + n_{j+1} + \ldots + n_{j+m} + \ldots + n_k$. Throughout the exposition of the theory and examples given in the next sections, it will be obvious that the group representations can be extended, starting from small genomic regions till cover whole chromosomes and, consequently, the whole genome, i.e., the set of all chromosomes.

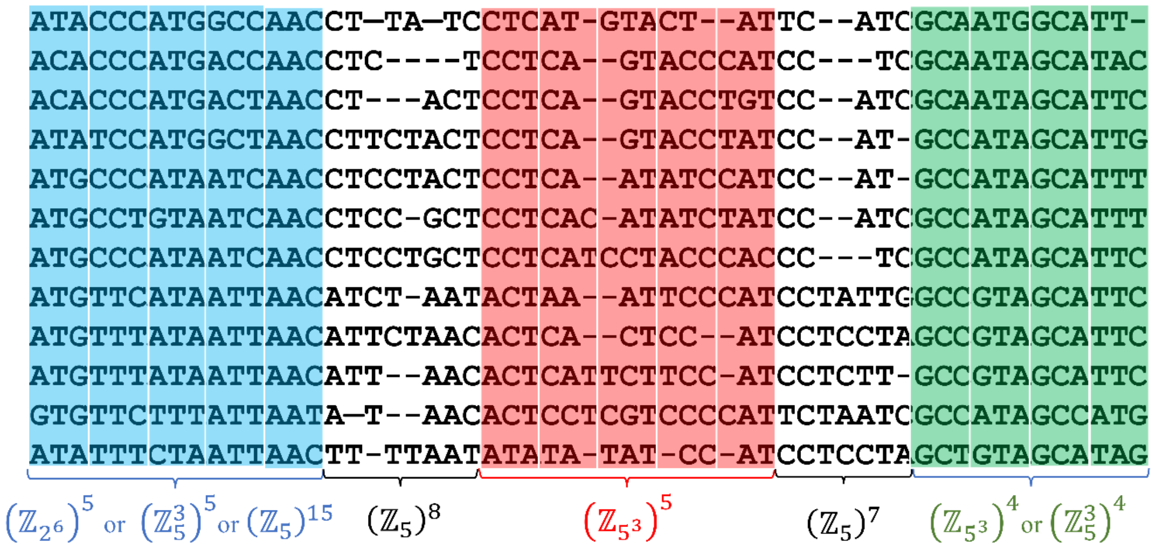
Let B_i $\left(i \in I = \{1, \ldots, n\}\right)$ be a family of subgroups of G, subject to the following two conditions:

1) $\sum B_i = G$. That is, B_i together generate G.

2) For every $i \in I$ and $i \neq j$: $B_i \cap \sum B_j = 0$.

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

Genomic DNA sequences from superior organisms 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 substitutions, insertions, and deletion (*indel*) mutations. As a result, the alignment of homolog genomic regions or 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., as the direct sum of cyclic group of the same *prime-power* order (Fig 3). As a result, each DNA sequence is represented as a *N*-dimensional vector with numerical coordinates representing bases and codons.



```
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--CTCC--ATCCTCCTAGCCGTAGCATTC
ATGTTTATAATTAACATT--AACACTCATTCTTCC-ATCCTCTT-GCCGTAGCATTC
GTGTTCTTTATTAATA—T--AACACTCCTCGTCCCCATTCTAATCGCCATAGCCATG
ATATTTCTAATTAACTT-TTAATATATA-TAT-CC-ATCCTCCTAGCTGTAGCATAG
```

$(\mathbb{Z}_{2^6})^5$ or $(\mathbb{Z}_5^3)^5$ or $(\mathbb{Z}_5)^{15}$
$(\mathbb{Z}_5)^8$
$(\mathbb{Z}_{5^3})^5$
$(\mathbb{Z}_5)^7$
$(\mathbb{Z}_{5^3})^4$ or $(\mathbb{Z}_5^3)^4$

Fig 3. An illustration of 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. A homocyclic group covering a MSA block corresponds to a sub-classification of the protein-coding region into subregions and, consequently, leading to a more accurate molecular taxonomy of species. In protein-coding regions cyclic groups $\mathbb{Z}_{2^6}$ and $\mathbb{Z}_5^3$ are appropriated to study exon regions, while $\mathbb{Z}_5$ for non-coding intron regions. As shown in section 1.4, the group representation leads us the analysis of the more frequent mutational events (represented as endomorphisms and translations) observable in genes from organismal populations.

An intuitive mathematical representation of a MSA is implicit in Fig 3, with the following observations:

a) Bases or codons can be represented as elements of an Abelian group defined on the set of bases or on the set of codons. In the second block (including gap symbol '-') each base from each sequence is represented as an element from the Abelian group defined on the set {A, C, G, T, D } where $D = '-'$, which is isomorphic to the Abelian p-group defined on the set $\mathbb{Z}_5$. The extended base triplets (including gaps symbol '-') from each sequence in the third aligned block are represented as elements from the Abelian p-group defined on the set of extended base-triplets (125 element, see SI Table 1) which is isomorphic to the Abelian group defined on the set $\mathbb{Z}_{5^3}$, and so on.

b) Every DNA sequence from the MSA and every subsequence on it can be represented as a *numerical vector* with element coordinates defined in an Abelian group. For practical computational purposes we take advantage of the group isomorphism to work with numerical representations of DNA bases and codons. For example, codons from the first aligned block (in blue) can be represented as elements from an Abelian group defined on the set of codons, which can be isomorphic to $\mathbb{Z}_{2^6}$ or to $\mathbb{Z}_5^3$. That is, since $\left(C_g,+\right) \cong \left(\mathbb{Z}_{2^6},+\right)$, the first five codons $\{\text{ATA, CCC, ATG, GCC, AAC}\} \in C_g$ from the first DNA sequence from Fig 3, can be represented by the vector of integers: $\{48, 21, 50, 25, 1\}$ where each coordinate is an element from group $\left(\mathbb{Z}_{2^6},+\right)$ (see Table 1 from reference [18]).

c) 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 p-groups. For example, the MSA from Fig 3 can be represented by the direct sum of five homocyclic Abelian p-groups:

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

Where the length of each region determines the number of cyclic p-groups in the corresponding homocyclic Abelian p-group $\mathbb{Z}_{p_1^{\alpha_1}}$ representing each region. For example, in Eq. 2 we have the homocyclic group: $\left(\mathbb{Z}_{5^3}\right)^4 = \oplus_{i=1}^{4}\mathbb{Z}_{5^3}$, which is a direct sum of 4 cyclic 5-groups $\left(\mathbb{Z}_{5^3},+\right) \cong \left(C_{g+},+\right)$. Since group G is the direct sum of homocyclic Abelian p-groups of different prime-order, we shall say that G is a heterocyclic group.

In more specific scenario, the MSA from Fig 3 can be represented by only one homocyclic Abelian 5-group:

$$G = (\mathbb{Z}_5)^{57} \qquad (3)$$

But this representation ignores the local variability detected by the MSA algorithm. Hence, preserving the highlighted features, the MSA can be represented as the direct sum of homocyclic 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 p-groups provides a useful compact representation of a MSA, for application purposes to genomics, we would also consider to use the concept of direct product (*cartesian sum or complete direct sums*) [30]. Next, let S be a set of Abelian cyclic groups identified in a MSA M of length N (i.e., every DNA sequence from M has N bases). Let ℓ_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 $\left(b_1,\ldots,b_n\right)$ where $b_i \in S_i$ $\left(i = 1,\ldots,n\right)$ and $n = \left|S\right|$. Let $\mathcal{G}_i$ be a group defined on the set of all elements $\left(0,\ldots,0,b_i,0,\ldots0\right)$ 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 $\left(b_1,\ldots,b_n\right)$ 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.:

$$\mathcal{G} = \otimes_i S_i = \oplus_i \mathcal{G}_i \tag{5}$$

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 [29,30]. By this theorem, every finite Abelian group G is isomorphic to a direct sum of cyclic groups of prime-power order of the form:

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

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 3 given by Eq. 2 correspond to a heterocyclic group that split into a direct sum of homocyclic Abelian 2-groups and 5-groups, each one of them split into the direct sum of cyclic p-groups with same order; while in Eqs. 3 and 4, the Abelian group G is decomposed into a direct sum of homocyclic Abelian 5-groups [29,30].

Notice that for a large enough genomic region of fixed length N we can build a *manifold of (a set of various) heterocyclic groups* S_i, where each one of them can have different decomposition into p-groups. The set S of all possible Abelian p-group representations S_i of a large genomic region of fixed length (having numerous different parts, elements, features, forms, etc.) that split into the direct sum of several heterocyclic groups G_k ($S_i = \oplus_{k=1}^{n} G_k$) shall be called a *heterocyclic-group manifold.* So, each genomic region can be characterized by means of their corresponding *heterocyclic-group manifold*.

3.1 Examples of genomic regions group representations

A group representation is particularly interesting for the analysis of DNA sequence motifs, which typically are highly conserved across the species. As suggested in Fig 3 and 4, there are 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 (TFs, Fig 4), which are identifiable applying bioinformatic algorithms like BLAST [31].

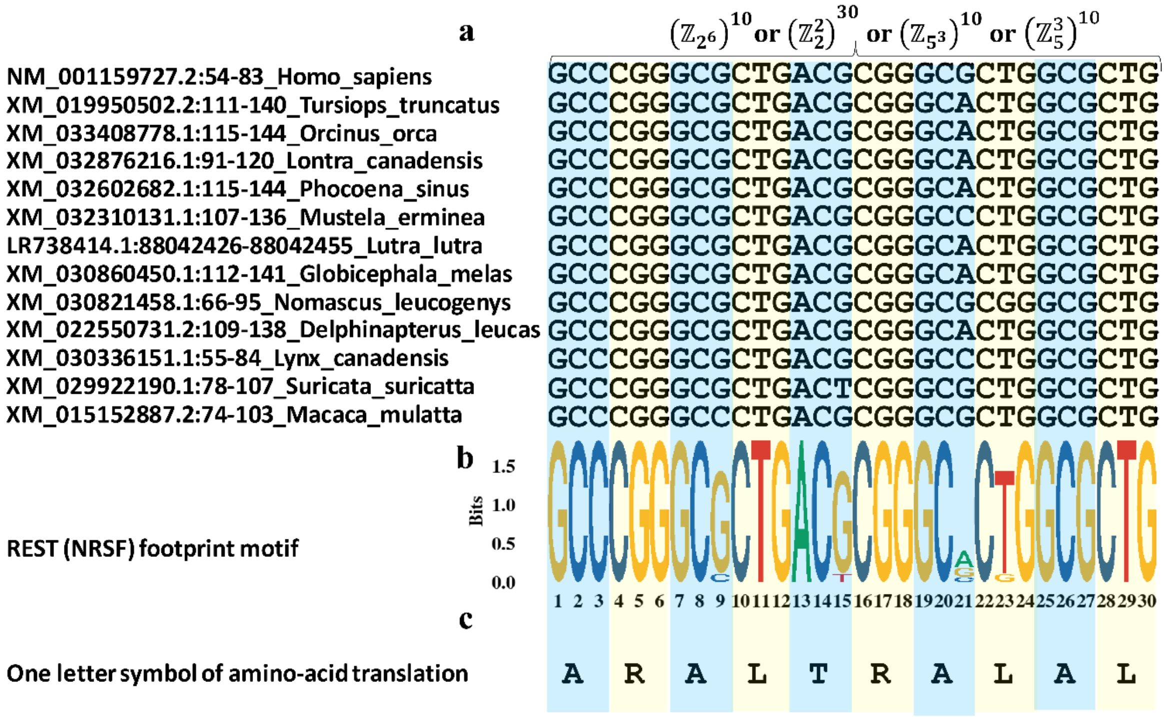

a
$\left(\mathbb{Z}_{2^6}\right)^{10}$ or $\left(\mathbb{Z}_2^2\right)^{30}$ or $\left(\mathbb{Z}_{5^3}\right)^{10}$ or $\left(\mathbb{Z}_5^3\right)^{10}$
NM_001159727.2:54-83_Homo_sapiens GCCCGGGCGCTGACGCGGGCGCTGGCGCTG
XM_019950502.2:111-140_Tursiops_truncatus GCCCGGGCGCTGACGCGGGCACTGGCGCTG
XM_033408778.1:115-144_Orcinus_orca 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
1.5
1.0
0.5
0.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
c
One letter symbol of amino-acid translation
A R A L T R A L A L

Fig 4. The DNA sequence motifs targeted by transcription factors usually integrate genomic building block across several mammal 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. **c**, Translation of the codon sequences using the one-letter symbol of the aminoacids.

The case of group representation on a TF binding motif is exemplified in Fig 4, 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. Herein, the case of double encoding called our attention, where the DNA sequence simultaneously encodes the information for transcription enhancer target motif and for a codon sequence (base-triplets) encoding for aminoacids. These types of double-coding regions are also called *duons* [32–34].

Four group representations for this exon subregion are suggested in the top of the Fig 4 (panel **a**). However, the MSA's sequence logo (panel **b**) suggests that this transcription factor binding-motif is a highly conserved codon sequence in mammals (with no indel mutations on it) and, in this case, the Abelian group $(C_g,+) \cong (\mathbb{Z}_{2^6},+)$ defined on the standard genetic code is the appropriated model to represent these motifs (Fig 4). The homocyclic group representation of conserved and biological relevant DNA sequence motifs, illustrated in Figs. 3 and 4, stablish the basis for the study of the molecular evolutionary process in the framework of group endomorphisms and automorphisms as suggested in [18,20] (section 1.4).

In Fig 5, two different protein-coding (gene) models from two different genome populations can lead to the same direct sum of Abelian *p*-groups and to the same final aminoacids sequence (protein).

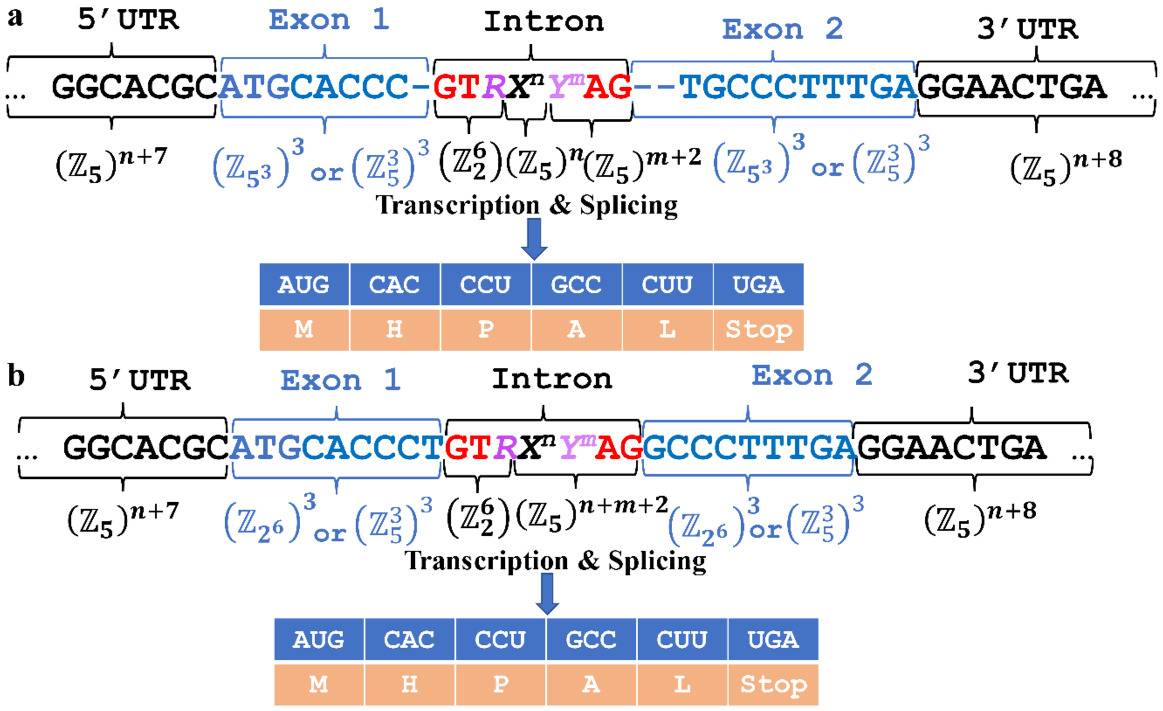

a
5′UTR
Exon 1
Intron
Exon 2
3′UTR
... GGCACGCATGCACCC-GTRX^nY^mAG--TGCCCTTTGAGGAACTGA ...
$(\mathbb{Z}_5)^{n+7}$
$(\mathbb{Z}_{5^3})^3$ or $(\mathbb{Z}_5^3)^3$
$(\mathbb{Z}_2^6)(\mathbb{Z}_5)^n(\mathbb{Z}_5)^{m+2}$
$(\mathbb{Z}_{5^3})^3$ or $(\mathbb{Z}_5^3)^3$
$(\mathbb{Z}_5)^{n+8}$
Transcription & Splicing
AUG CAC CCU GCC CUU UGA
M H P A L Stop
b
5′UTR
Exon 1
Intron
Exon 2
3′UTR
... GGCACGCATGCACCCTGTRX^nY^mAGGCCCTTTGAGGAACTGA ...
$(\mathbb{Z}_5)^{n+7}$
$(\mathbb{Z}_{2^6})^3$ or $(\mathbb{Z}_5^3)^3$
$(\mathbb{Z}_2^6)(\mathbb{Z}_5)^{n+m+2}$
$(\mathbb{Z}_{2^6})^3$ or $(\mathbb{Z}_5^3)^3$
$(\mathbb{Z}_5)^{n+8}$
Transcription & Splicing
AUG CAC CCU GCC CUU UGA
M H P A L Stop

Fig 5. 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 (GUR) and acceptor (Y^mAG) 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 gene models, from panels **a** and **b**, share a common group representation as direct sum of Abelian 5-groups.

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**, share a common direct sum of Abelian 2-groups and 5-groups: $(\mathbb{Z}_5)^{n+7} \oplus (\mathbb{Z}_5^3)^3 \oplus (\mathbb{Z}_2^6) \oplus (\mathbb{Z}_5)^{n+m+2} \oplus (\mathbb{Z}_5^3)^3 \oplus (\mathbb{Z}_5)^{n+8}$. The analysis of theses gene models suggests that *DNA sequences sharing a common group representation as direct sum of Abelian p-groups would carry the same or similar, or close related biological information*. However, it does not imply that the architecture of these protein-coding regions is the same. The gene model in panel **b** permits the direct sum representation: $(\mathbb{Z}_5)^{n+7} \oplus (\mathbb{Z}_{2^6})^3 \oplus (\mathbb{Z}_2^6) \oplus (\mathbb{Z}_5)^{n+m+2} \oplus (\mathbb{Z}_{2^6})^3 \oplus (\mathbb{Z}_5)^{n+8}$, which is no possible for the gene model from panel **a**. That is, the *heterocyclic-group manifold* from the gene model in panel **a** is different from the one in panel **b**. The difference of group representation just captures the obvious fact that these gene models are different and, consequently, their gene architectures are different.

At this point we shall introduce the concept of *equivalent class of genomic region*. We shall say that two genomic regions belong to same *equivalent class of genomic region* if they hold the same heterocyclic-group manifold (and, consequently, they hold same architecture). Under this definition, the region architecture of the protein-coding regions from Fig 5**a** and **b** are not equivalent. The concept of *equivalent class of genomic region* is relevant for further applications of the group representation on the taxonomy study of organismal populations.

Taxonomy is the study of the scientific classification of biological organisms into groups based on shared characteristics. Mathematically, this is a way to split biological organisms into classes of equivalences. Numerical taxonomy is a well-established application of multivariate statistics on the analysis of plant germplasm banks. The group representations of genomic regions will lead to a higher accuracy in the taxonomy study of organismal populations.

No matter how complex a genomic region might be, it has an Abelian group representation. A further application of group theory would unveil more specific decomposition of small genomic regions into Abelian groups. For example, the set of base-triplets found in a typical sequence motif targeted by the spliceosome donor, GT*R* (Figs. 5 and 6), is in the vertical line GT*Z* (GU*Z*) of the vertical plane *X*T*Z* (*X*U*Z*) from the cube ACGU shown in Fig 2 (see also SI Fig 3).

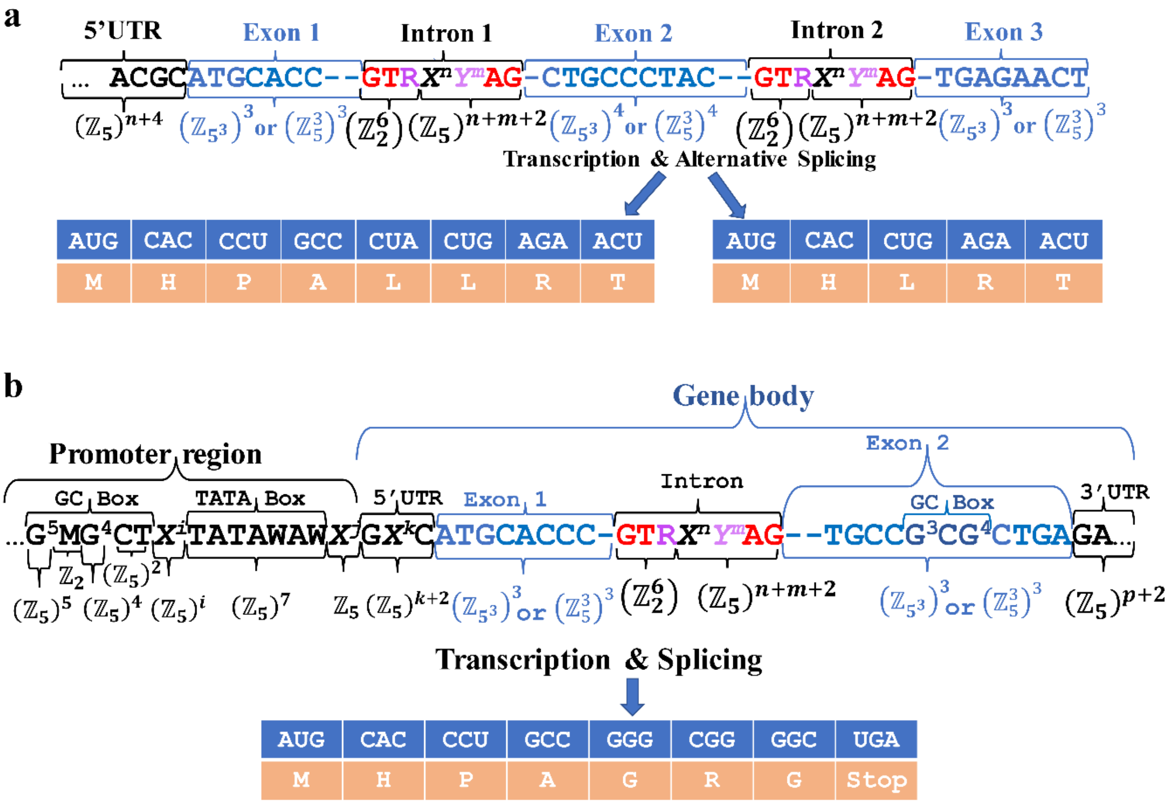

a
5'UTR
Exon 1
Intron 1
Exon 2
Intron 2
Exon 3
... ACGCATGCACC--GTRX^nY^mAG-CTGCCCTAC--GTRX^nY^mAG-TGAGAACT
Transcription & Alternative Splicing
AUG CAC CCU GCC CUA CUG AGA ACU
M H P A L L R T
AUG CAC CUG AGA ACU
M H L R T
b
Gene body
Promoter region
Exon 2
GC Box
TATA Box
5'UTR
Exon 1
Intron
GC Box
3'UTR
...G^5MG^4CTX^iTATAWAWX^7GX^kCATGCACCC-GTRX^nY^mAG--TGCCG^3CG^4CTGAGA...
Transcription & Splicing
AUG CAC CCU GCC GGG CGG GGC UGA
M H P A G R G Stop

Fig 6. 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). In this group representation, the spliceosome donor GT*R* can be represented by the elements from a quotient group (see main text).

Since purine bases (R: A and G) are the only accepted variants at the third codon position, it is convenient to model these base-triples with the group defined on the cube AGCU [11] (SI Fig 3). Next, following analogous reasoning as in [19], it turns out that the set of base-triplets GT*R* is a coset from the quotient group $\left(C_G,+\right)/G_{\mathrm{AAG}}$, where here $\left(C_G,+\right)\cong\left(\mathbb{Z}_2^6,+\right)$ is the additive group from the genetic code Galois field $GF(64)$ reported in reference [35] and $G_{\mathrm{AAG}}=\left(\left\{\mathrm{AAA,AAG}\right\},+\right)$ is a subgroup from the Klein four group defined on the set $\left\{\mathrm{AAA,AAG,AAC,AAU}\right\}$ (see operation table in the SI Table 2), i.e., $\mathrm{GT}R=\mathrm{GTA}+G_{\mathrm{AAG}}$ (SI Fig 3).

There exists strong evolutionary pressure on splicing donor site to keep the base-triplet GTR in the vertical line GTZ (GUZ) vertical line (coset). As shown in the clinical report [36] mutational variants, located in different cube's vertical lines (different cosets, SI Fig 3) GCZ and CTZ (CUZ), within intron 3 have led to four aberrant RNAs transcripts that causes rare X-chromosome-linked congenital deafness. As will be shown below (in section 3.1) the strong connection between DNA sequences and non-disrupting mutational events is mathematically (and accurately) modeled by the strong relationship between a group representation and the endomorphism ring on it.

An example considering changes on the gene-body reading frames as those observed in alternative splicing is shown in Fig 6. Gene-bodies with annotated alternative splicing can easily be represented by any of the groups $\left(\mathbb{Z}_5^3\right)^n$ or $\left(\mathbb{Z}_{5^3}\right)^n$ (Fig 6**a**). The splicing can include enhancer regions as well (Fig 6**b**) [37]. Enhancers are key regulator of differential gene expression programs.

As commented in the introduction, cytosine DNA methylation is implicitly included in extended base-triple group representation. Typically, the analysis of methylome data 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 adenine and cytosine ($D=C^m$), since only epigenetic changes are evaluated.

Concrete examples of adenine in bacteria linked to the regulation of pyelonephritis-associated pilus (pap) expression by DNA methylation on the *Escherichia coli* operon (locus X14471) and cytosine methylation in two (humans) genes from patients with pediatric acute lymphoblastic leukemia (PALL) are presented in Fig 7. On protein-coding regions methylation change can be analyzed on the homocyclic groups composed by the cyclic group $\mathbb{Z}_5^3$ or $\mathbb{Z}_{5^3}$ (Fig 7c and **d**). Notice that adenine methylation is found in humans as well and, usually, it plays a very specific regulatory role [38,39].

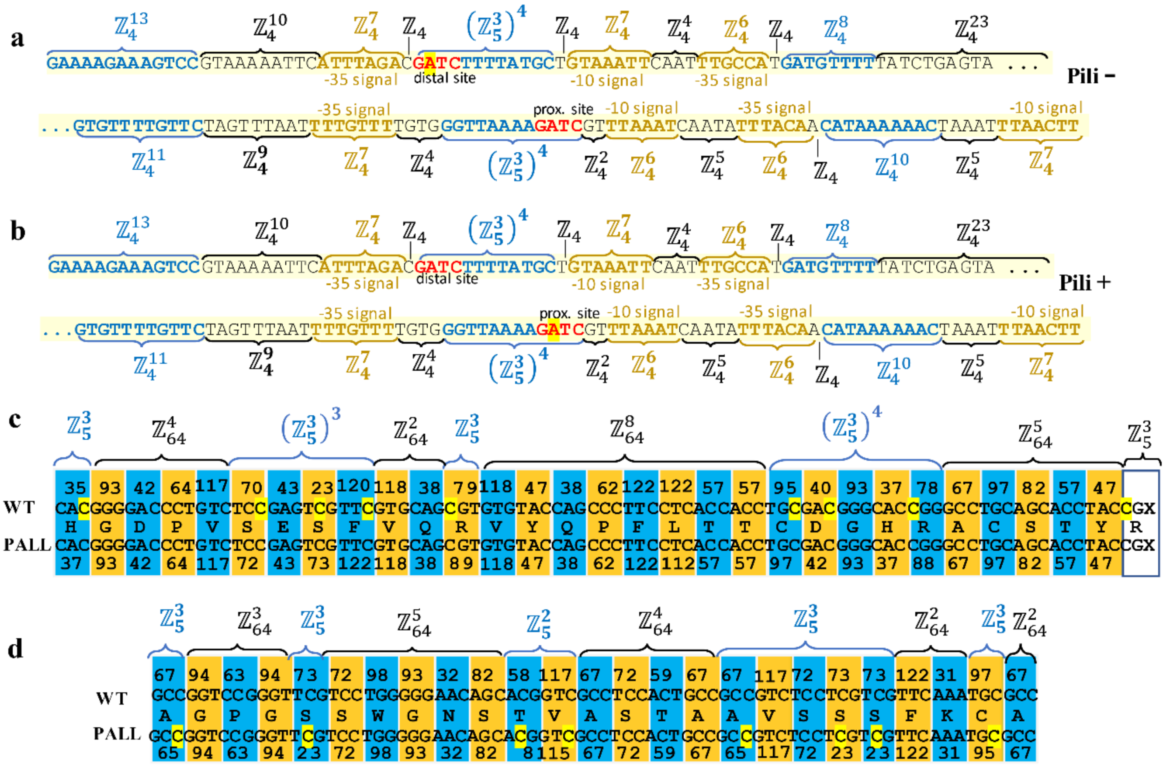

a
b
c
d
-35 signal
distal site
-10 signal
prox. site
Pili −
Pili +
WT
PALL
GAAAAGAAAGTCCGTAAAAATTCATTTAGACGATCTTTTATGCTGTAAATTCAATTTGCCATGATGTTTTTATCTGAGTA . . .
. . . GTGTTTTGTTCTAGTTTAATTTTGTTTTGTGGGTTAAAAGATCGTTTAAATCAATATTTACAACATAAAAAACTAAATTTAACTT

Fig 7. Vector representation of differentially methylated gene regions. **a** and **b**, regulation of pyelonephritis-associated pilus (pap) expression by DNA methylation on the Escherichia coli operon (locus X14471). **c** and **d**, exons regions from genes EGEL7 and P2RY1 from patients with pediatric acute lymphoblastic leukemia (PALL). In panel **a**, two 5'-GATC-3' DNA adenine methyltransferase (Dam) methylation sites in the middle of each set of the leucine-responsive regulatory protein (Lrp) binding sites (in blue). In the inactive state, panel **b**, a Lrp octamer is bound to the three proximal Lrp 3' sites, while the GATC^{dist} site in Lrp site 5 is fully methylated, and the system remains in phase OFF (Pili -) with regard to pilus expression. In the active state, the adenine from the GATC^{prox} is methylated permitting to bend the DNA to recruit CRP to activate transcription of papBA genes (Pili +). Pap pili are multisubunit fibers essential for the attachment of uropathogenic Escherichia coli to the kidney (see [40]). In panel **c**, a segment of exon-6 from gene EGFL7 located at chromosome 9: 139,563,008-139,563,124 is shown. On average, this gene is hypo-methylated in the control group with respect to PALL group. **d**. 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 [41]. The integer number at the top and bottom of panel **c** and **d** stand for the codon coordinates in $\mathbb{Z}_{5^3}$ (see SI Table 1).

It is obvious that the MSA from a whole genome 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 homocyclic 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 homocyclic groups of prime-power order. Hence, results lead us to the representation of genomic regions from organismal populations from the same species or close related species (as suggested in Fig 3 to 7) by means of direct sum of their group representation into Abelian cyclic groups. A general illustration of this modelling would be, for example:

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

That is, Eq. 7 expresses that any large enough genomic region can be represented as direct sum of homocyclic Abelian groups of prime-power order. In other words, the fundamental theorem of Abelian finite groups (FTAG) 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 homocyclic Abelian groups of prime-power order.

The proof of this theorem is self-evident across the discussion and examples presented here. Basically, group representations of the genetic code lead to group representations of local genomic domains in terms of cyclic groups of prime-power order, for example, $(C_g,+) \cong (\mathbb{Z}_{2^6},+)$, $(C_{G+},+) \cong (\mathbb{Z}_5^3,+)$ or $(C_{g+},+) \cong (\mathbb{Z}_{5^3},+)$, till covering the whole genome. As for any finite Abelian group, the Abelian group representation of genome populations can be expressed in terms of a direct sum of Abelian homocyclic groups of prime-power order. Any new discovering on the annotation of a given genome population will only split an Abelian group, already defined on some genomic domain/region, into the direct sum of Abelian subgroups ∎.

The application of the FTAG in terms of the group representation of genomic regions *G*, as given in Eq. 7, establishes the basis to the study the molecular evolutionary process in terms of endomorphisms. That is, fixed mutational events in the organismal population can be modeled as homomorphism: endomorphisms and automorphisms, all elements of the endomorphism ring $\mathfrak{R}(G)$ on G (see next section). In the context of comparative evolutionary genomics, the analysis of the endomorphism ring $\mathfrak{R}(G)$ is an intermediate step for the further application of methods from Category theory, which has the potential to unveil unsuspected features of the genome architecture, hard to be inferred from the direct experimentation.

3.2 The endomorphism ring

A biologically relevant application of the theory presented here relies on the fact that if a finite group G is written as a direct sum of subgroups G_i, as given in Eq. 7, then endomorphism ring $End(G)$ is isomorphic to the ring matrices (A_{ij}), where $A_{ij} \in Homo(G_i, G_j)$ (homomorphism between G_i and G_j), with the usual matrix operations [30]. In the case of genomic regions from the species or closed related genomic regions from distinct species, the endomorphism that transform the DNA aligned sequence α into β (α, $\beta \in G$) is represented by a matrix with only non-zero elements in the principal diagonal. These diagonal elements are sub-matrices $A_{ii} \in End(G_i)$ or $A_{ii} \in Aut(G_i)$. In other words, mutational events fixed in gene/genome populations can be quantitatively described as endomorphisms and automorphisms.

In the Abelian p-group defined on $\mathbb{Z}_{p_i^{\alpha_i}}$, the endomorphisms $\eta_i \in End\left(\mathbb{Z}_{p_i^{\alpha_i}}\right)$ are described as functions $f(x) = k\,x \bmod p_i^{\alpha_i}$, where k and x are elements from the set of integers modulo $p_i^{\alpha_i}$. For example, in the cube ACGT the sequence ATACCCATGGCCAAC (blue block in Fig. 3) represented by the vector $(48, 21, 50, 25, 1) \in \left(\mathbb{Z}_{2^6}\right)^5$ is transformed into the sequence

ACACCCATGACCAAC, represented by the vector $(16,21,50,17,1)\in\mathbb{Z}_{2^6}$, by the automorphism:

$$\begin{bmatrix}3&0&0&0&0\\0&1&0&0&0\\0&0&1&0&0\\0&0&0&57&0\\0&0&0&0&1\end{bmatrix}\text{, i.e.: }(48,21,50,25,1)\begin{bmatrix}3&0&0&0&0\\0&1&0&0&0\\0&0&1&0&0\\0&0&0&57&0\\0&0&0&0&1\end{bmatrix}\bmod 64=\begin{pmatrix}16\\21\\50\\17\\1\end{pmatrix}.$$

Now, it is not difficult to realize that the set of all endomorphisms $\eta_i\in End\left(\mathbb{Z}_{p_i^{\alpha_i}}\right)$ hold the ring axioms mentioned in the Introduction. That is, the set of all endomorphisms $\eta_i\in End\left(\mathbb{Z}_{p_i^{\alpha_i}}\right)$ forms a ring on $\mathbb{Z}_{p_i^{\alpha_i}}$ that we shall denote as $\mathfrak{R}\left(\mathbb{Z}_{p_i^{\alpha_i}}\right)$.

As shown in reference [30], if $G=G_1\oplus G_2\ldots\oplus G_n$ *is a direct decomposition with fully invariant summands*, then :

$$End(G)=End(G_1)\oplus End(G_2)\ldots\oplus End(G_n) \tag{8}$$

In this modeling, mutational events are represented as endomorphisms $\eta_i\in End\left(\mathbb{Z}_{p_i^{\alpha_i}}\right)$ on $\mathbb{Z}_{p_i^{\alpha_i}}$. This fact permits the study of the genome architecture through the study of the evolutionary (mutational) process in a genome population. Moreover, the decomposition of the endomorphism ring into subgroups, quotient groups, and cosets can lead to a deterministic algebraic taxonomy of the species based on their genome architecture, which is not limited by our current biological knowledge.

Particularly relevant for the evolutionary comparative genomics is Baer-Kaplansky theorem: *If G and H are p-groups such that* $\mathfrak{R}(G)\cong\mathfrak{R}(H)$, then $G\cong H$ ([29,42]). That is, two Abelian finite groups are isomorphic if, and only if, their endomorphism rings are isomorphic [42]. In other words, genomic regions experiencing mutational events representable by isomorphic rings are algebraically represented by isomorphic Abelian groups and, consequently, have similar genome architecture.

Application of Baer-Kaplansky theorem implies that two gene-body regions encoding exactly for the same polypeptide but with different region architecture (Fig 5) are under different evolutionary pressure. That is, if the group representations of two gene-body regions are not isomorphic, then their endomorphism rings are not isomorphic either and, consequently, they will be under different evolutionary pressure, experiencing different subsets of mutational events, which are represented as endomorphisms from their corresponding endomorphism ring. This scenario is typically found in some isoforms, which are proteins that are similar to each other and perform similar roles within cells [43]. This is the case where two or more closely related genes are responsible for the same translated protein, illustrated in Fig 5. They can be simply duplicated, or paralogous genes, where both paralogs can remain similar (paralog isoforms) if an increased production of the protein is advantageous or if a dosage balance occurs in conjunction with other gene products or where different transcripts can lead to different subcellular localization [44].

A screening of mutational events on subsets of aligned genes suggests that the decomposition of protein-coding regions is tractable, conforming Eq. 8. Results with the alignments of several protein-coding regions are shown in Fig 8. In this example, we searched for automorphisms on the *groups of dual cubes* [11]: ACGT – TGCA and CATG – GTAC on $\mathbb{Z}_{2^6}$, which comprise four of the 24 possible algebraic representations of the standard genetic code [17] isomorphic to $\mathbb{Z}_{2^6}$.

The analysis of the frequency of mutational events (automorphisms, COVID: human *vs* bat strains) by mutation types is shown in Fig 8**a**. Results are consistent with the well-known observation highlighted by Crick: *the highest mutational rate is found in the third base of the codon, followed by the first base, and the lowest rate is found in the second one* [45]. However, estimations on different gene sets suggest that the evolutionary pressure on each codon position depends on the physicochemical properties (annotated according to IUPAC nomenclature [36]) of DNA bases. For example, in Fig 8**a** pyrimidine (Y) transitions on the third codon position (HHY) are, by far, the most frequent observed mutational events. While, in BRCA1 gene (SI Fig 2), the frequency of purine (HHR) transitions is followed by pyrimidine (HHY) transitions.

The analysis on the pairwise alignment of protein-coding regions of SARS and Bat SARS-like coronaviruses is presented in Fig 8**b** an **c**. Most of the mutational events distinguishing human SARS from Bat SARS-like coronaviruses can be described by automorphism on cube ACGT. This observation was confirmed in primate somatic cytochrome c (Fig 8**c**) and BRCA1 DNA repair gene

(Fig 8**d**). Since automorphisms transform the null element (gap-triplet DDD/---) into itself, insertion-deletion mutational events cannot be described by automorphisms but as translations on the groups (denoted as *Trnl* in Fig 8). The representation of conserved genomic regions with homocyclic *p*-group is straightforward. However, their frequency in the genome architecture exponentially decreases with the size of the region (Fig 8**f** and SI Fig 4).

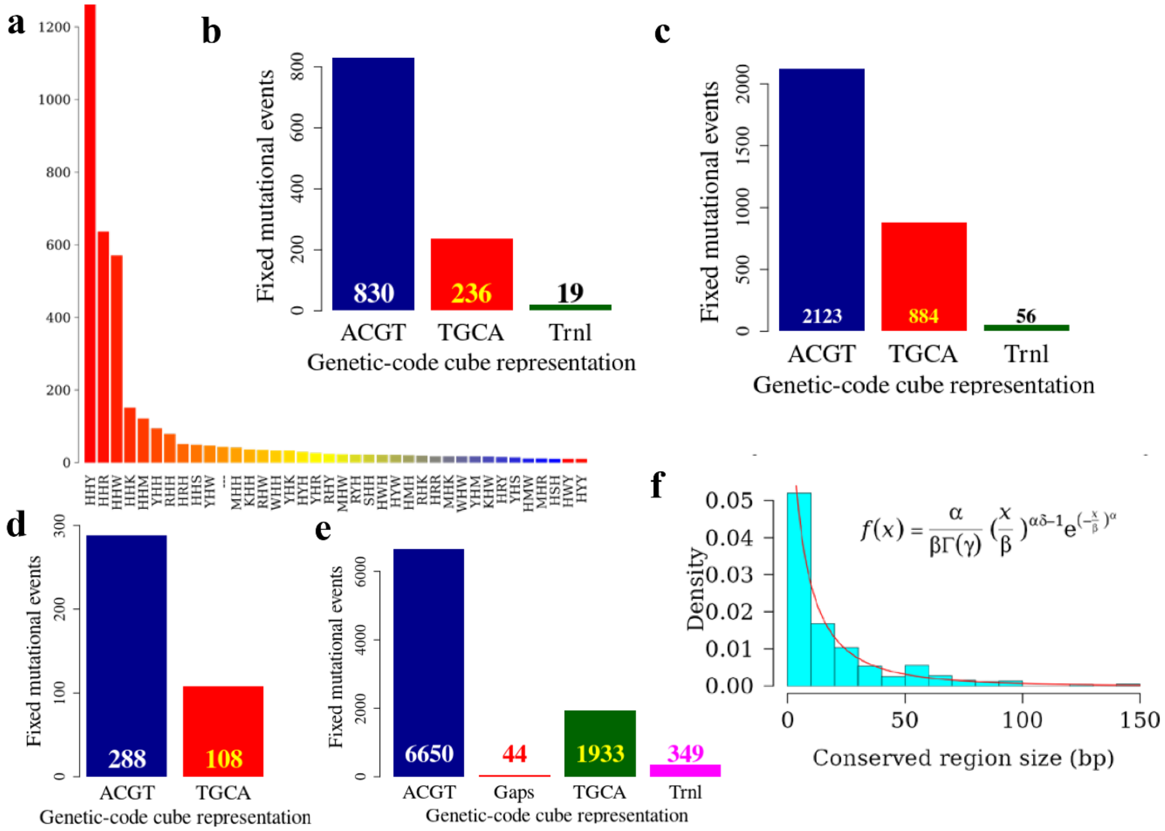


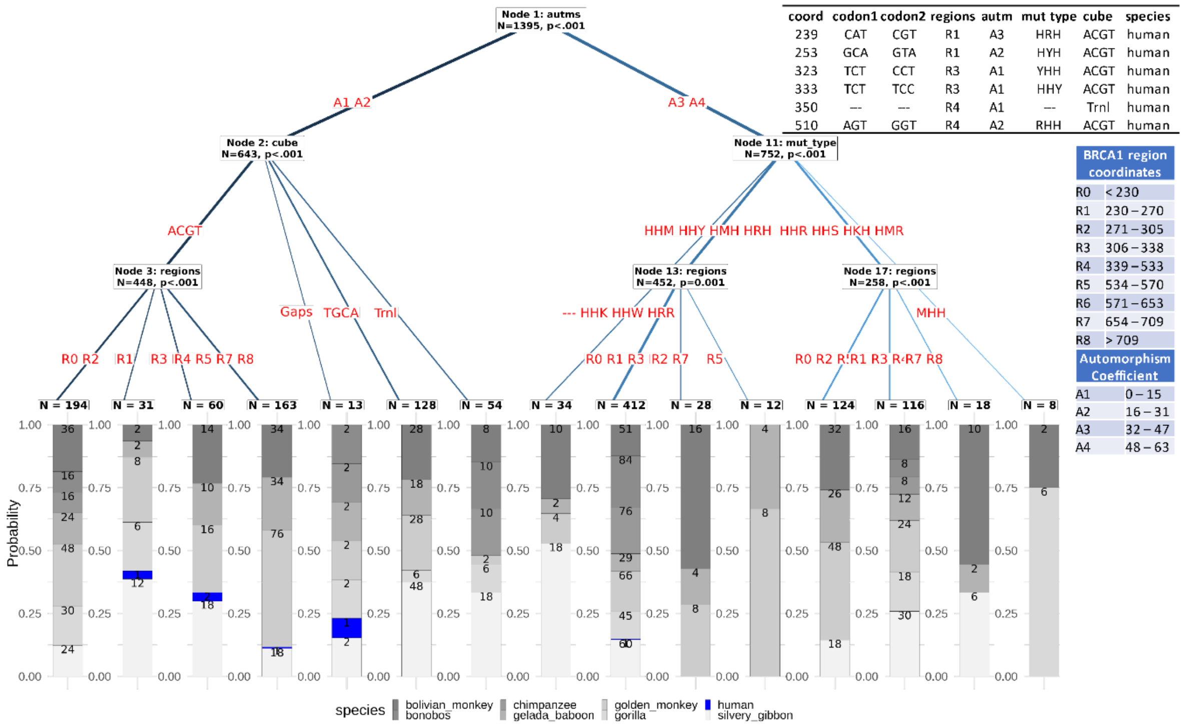
Fig 8. Analysis of mutational events in terms of automorphisms on DNA protein-coding regions represented as homocyclic groups on $\mathbb{Z}_{64}$. In the Abelian group defined on $\mathbb{Z}_{64}$, automorphisms are described as functions $f(x) = kx \bmod 64$, where k and x are elements from the set of integers modulo 64. **a**, Frequency of mutational events (automorphisms) according to their mutation type. That is, every single base mutational event across the MSA was classified according IUPAC nomenclature [46]: 1) According to the number of hydrogen bonds (on DNA/RNA double helix): strong S={C, G} (three hydrogen bonds) and weak W={A, U} (two hydrogen bonds). According to the chemical type: purines R= {A, G} and pyrimidines Y= {C, U}. 3). According to the presence of amino or keto groups on the base rings: amino M= {C, A} and keto K= {G, T}. Constant (hold) base positions were labeled with letter H. So, codon positions labeled as HKH means that the first and third bases remains constant and mutational events between bases G and T were found in the MSA. **b** and **c**, Bar plots showing the frequency of automorphisms found on the *group of dual cubes* (see [11]): ACGT – TGCA and CATG – GTAC on $\mathbb{Z}_{64}$ between SARS coronavirus GZ02 and bat SARS-like coronaviruses: **a**, isolate Rs7327 (GenBank: KY417151.1, protein-coding regions) and **c**, isolate bat-SL-CoVZC45 (GenBank: MG772933.1:265-1345513455-21542, nonstructural polyprotein). **d**, frequency of automorphisms between human somatic cytochrome c and other nine primates (monkeys). **e**, frequency of automorphisms between human BRCA1 DNA repair gene and other seven primates (see Material and

Method section). **f**, Distribution of the conserved COVID-19 genomic regions according to their size. The graphics result from the analysis SARS coronavirus GZ02 versus the two mentioned bat strains. The best fitted probability distribution turned out to be the generalized gamma distribution.

Next, under the assumption that Eq. 8 holds, different protein-coding regions must experience "*preference*" for specific type of automorphisms. To illustrating the concept, an analysis based on the application of Theorem 1 and Eq. 8 on gene/genome population studies, an application of decision tree algorithms was conducted on primate BRCA1 genes. Results for the analysis with Chi-squared Automated Interaction Detection (CHAID) is presented in Fig 9. It is important to keep in mind that this is only an illustrative example with small sample size, and that definite conclusions related to BRCA1 genes can only be derived with larger sample size from humans and non-human primate sequences. In this algorithmic approach, for each compound category consisting of three or more of the original categories, the algorithm finds the most significant binary split for a node (split-variable) based on a chi-squared test [47].

For a given MSA of protein-coding regions, the resulting decision tree leads to stochastic-deterministic logical rules (propositions) permitting a probabilistic estimation of the best model approach holding Eq. 8. For example, since only one mutational event human-to-human from class A3 is reported in the right side of the tree (Fig 9), with high probability the proposition: "(A4 $\vee$ (A3 $\wedge$ $\neg$ HRH) $\rightarrow$ $\neg$ human" is true. That is, with high probability only non-humans hold the last rule. Due to graphic printing limitations not all tree details are shown in Fig 9 (calculations details are given in the tutorials links provided at SI).

Results shown in Fig 9 are only for the purpose to illustrate the application of the theory, since for the sake of visualization and simplicity, were limited to small sample data set and to the application of a relatively "modest" (unsupervised) machine-learning approach which, however, is sufficient to illustrate the concepts. Next, let us suppose that the decision tree from Fig 9 holds on a large enough sample-size (to minimize the classification error) of primate BRCA1-gene populations. Then, with high probability the logical rule: "A1 $\wedge$ R3 $\wedge$ (YHH $\vee$ HHY) $\rightarrow$ human" is true. That is, with high probability transitions mutations (T ↔ C) on region R3 from BRCA1 gene (specifically at positions 323 and 333, Fig 9) in the first and third codon positions, represented by automorphisms with coefficient between 0 and 15, are not observed in primates other than humans.



coord	codon1	codon2	regions	autm	mut type	cube	species
239	CAT	CGT	R1	A3	HRH	ACGT	human
253	GCA	GTA	R1	A2	HYH	ACGT	human
323	TCT	CCT	R3	A1	YHH	ACGT	human
333	TCT	TCC	R3	A1	HHY	ACGT	human
350	---	---	R4	A1	---	Trnl	human
510	AGT	GGT	R4	A2	RHH	ACGT	human

BRCA1 region coordinates	
R0	< 230
R1	230 – 270
R2	271 – 305
R3	306 – 338
R4	339 – 533
R5	534 – 570
R6	571 – 653
R7	654 – 709
R8	> 709

Automorphism Coefficient	
A1	0 – 15
A2	16 – 31
A3	32 – 47
A4	48 – 63

Fig 9. Decision tree based on automorphisms estimated on primate BRCA1 genes. Symbols R0 to R8 denote the protein regions as given in UniProt plus inter regions segments (see https://www.uniprot.org/uniprot/P38398#family_and_domains). Only regions experiencing fixed mutational events are included in the analysis. The range of automorphism coefficients k ($f(x) = k\,x$ mod 64) are denoted after the isomorphism between the genetic-code cyclic group defined in the set of codons and the Abelian group defined on $\mathbb{Z}_{64}$. For the sake of graphic comprehension, the coordinates of human-to-human mutations were added. Every branch (path) from the top to the leaf node is equivalent to a stochastic-determinist logical rule defining the automorphism preference for each protein region in the subset of analyzed primate BRCA1 genes. For example, with high probability the rule: "(A4 ∨ (A3 ∧ ¬ R1)) → ¬ human" is true (see Supporting Information).

Obviously, the predictive power of the stochastic rules depends on the size of the samples from the populations under scrutiny. A larger data set including 41 variants of the BRCA1 gene and a rough estimation of the (encoded) *mutational cost* given in the term of a quasichemical energy of aminoacid interactions in an average buried environment [11,48] (data included in the GenomAutomorphism R package [28]) allow reach more robust rules after the application of decision tree algorithms. Likewise, an estimation of *mutational cost* can be given in terms of distances between aminoacids based on codon distances defined on a specific genetic-code cube model or on a combination of two models [11,49]. Examples of stochastic some mutational rules are given in Table 1.

Table 1. Examples of stochastic mutational rules found in aligned DNA sequences from primate BRCA1 genes.

Mutational cost (MC)	**Stochastic Rule**[3]
Aminoacid contact potential[1]	MC(0.03) $\rightarrow \neg$ human
	MC(-0.47) $\wedge$ R4 $\wedge$ A4 $\rightarrow$ human
	MC(-0.47) $\wedge$ (R0 $\vee$ R0. $\vee$ R3 $\vee$ R5) $\rightarrow$ bonobos
	MC(0.08) $\rightarrow$ bolivian_monkey
Aminoacid distance based on genetic-code codon distances[2]	MC(1.34) $\wedge$ R0 $\rightarrow \neg$ human
	MC(1.36) $\rightarrow$ gorilla
	MC(0.28) $\wedge$ R4 $\wedge$ A4 $\rightarrow$ human
	(MC(0.12) $\vee$ MC(0.12)) $\wedge$ (R1 $\vee$ R5) $\rightarrow$ silvery gibbon
	MC(0.26) $\wedge$ (A1 $\vee$ A2) $\wedge \neg$ R4 $\wedge \neg$ HHW $\rightarrow$ human
	MC(0.99) $\wedge$ HHS $\wedge$ (R7 $\vee$ R4) $\rightarrow$ golden monkey

[1]Aminoacid contact potentials are given in reference [48]. [2]Aminoacid distance based on the codon distance are given in reference [49] and applied (together with the concept of encoded mutational cost) in reference [11]. 3 The decision trees using CHAID algorithm are given in the Supporting Information (also available in the tutorials at https://genomaths.github.io/genomautomorphism).

Our results provides supporting evidence to the previous finding reported in [11] about that the selection of the genetic-code cube model cannot be arbitrary, since the automorphisms and the estimation of mutational costs (as defined in [11]) on different local DNA protein-coding regions shows clear "preference" for specific models. Obviously, the mathematical model is only a tool (a representation of the physicochemical relationships given between molecules) applied to uncovering the existence of specific evolutionary constraints.

3.3 Future theoretical developments

In this section we want to highlight a direction of future theoretical development. A full coverage of this topic is out of the limits of the current work. Nevertheless, a sketch on a future direction is presented here. Our goal will be the description of mutational process on protein-coding regions in terms of homomorphisms of different algebraic structures.

Genomic regions represented as an Abelian group decomposable into homocyclic Abelian *p*-groups, e.g. $\overset{n\ times}{\mathbb{Z}_{2^6} \oplus \ldots \oplus \mathbb{Z}_{2^6}}$, can be studied as *R*-algebras [18], which in particular is a *R*-module and after considering only the sum operation of the ring $\mathbb{Z}_{2^6}$, it is also a *G*-module. Recall that our modeling just takes advantage of the group isomorphism: $(\mathbb{Z}_{64},+) \cong (C_g,+)$ (for the sake of simplicity we are using the same sum operation symbol in both groups, $\mathbb{Z}_{64}$ and C_g). Thus, the $\mathbb{Z}_{64}$-algebra of the group $S = \left(\left(C_g \right)^n, + \right) = \left(C_g, + \right) \overset{n\ times}{\oplus \ldots \oplus} \left(C_g, + \right)$ over the ring $\mathbb{Z}_{64}$ can be defined [18].

In our current case (considering the codon coordinate level), we are interested on heterocycle groups $S = \oplus G_i$ of C_g and C_{g+} $(G_i \in \{C_g, C_{g+}\})$, as suggested in Fig. 1, which permits the analysis of multiple sequence alignments including insertion-deletion (indel) mutations. It is not hard to notice that the collection of all the R-Module of groups S over the ring $R = \otimes R_i$, $(R_i \in \{\mathbb{Z}_{64}, \mathbb{Z}_{125}\})$ together with R-Module homomorphisms conform to a category of ***R*-Modules**, also denoted as ***R*-Mod**. Let $\mathcal{C}_N$ be the category **Ab** with the Abelian groups of the DNA sequences of length $\leq N$ as objects and group homomorphisms as morphisms (see Appendix A). Fredy's theorem states that every Abelian category is a subcategory of some category of modules over a ring [50]. Mitchell has reinforced Fredy's result, proving that every Abelian category is a full subcategory of a category of modules over a ring [51].

At codon coordinate level, the group defined on the set of codon is a subgroup of the group defined on the set of extended base-triplets ($C_g \subset C_{g+}$) and the $\mathbb{Z}_{125}$-Module of group C_g is a submodule of the $\mathbb{Z}_{125}$-Module of group C_{g+} over the ring $\mathbb{Z}_{125}$. The triplet of gaps '---' corresponds to the identity element of group C_{g+}, which is mapped into $0 \in \mathbb{Z}_{5^3}$ by $Hom\left(C_{g+}, \mathbb{Z}_{5^3} \right)$. A homomorphism always maps the identity element from the domain of group, say $\mathbf{0}_{C_g}$, into the identity element from the codomain $\mathbf{0}_{C_{g+}}$, which in C_{g+} is $0_{C_{g+}} =$ '- - -'.

The following example illustrates a possible sequence of attainable analytical steps with concrete computational biology application. Let A = GACAGAGCAGTATTAGCTTCACAC and B = GAAAACGTATTATCAAAG DNA sequence segments represented as elements from the groups: $G_A = C_g^{\mathrm{ACGT}} \oplus C_g^{\mathrm{TGCA}} \oplus (C_g^{\mathrm{ACGT}})^6$ and $G_B = C_g^{\mathrm{ACGT}} \oplus C_g^{\mathrm{TGCA}} \oplus (C_g^{\mathrm{ACGT}})^4$, respectively, where C_g^X is the Abelian p-group defined on the set of 64 codons and base orders (cubes): $X = \{\mathrm{ACGT}, \mathrm{TGCA}\}$.

Groups G_A and G_B are elements of the **Ab** category $\mathcal{C}_N$ defined on the collection of heterocyclic group $\left(C_g^X\right)^N$ defined on the set of DNA sequences (of codons) with length $N \leq 8$.

Since the triplet of gaps cannot be arbitrary allocated in the sequence, the alignment of DNA sequence is an essential step required for the application of this modeling preserving the biological meaning. The pairwise alignment of the corresponding aminoacid sequences from A and B yields: $\begin{matrix}\texttt{DRAVLASQ}\\ \texttt{EN-VL-SN}\end{matrix}$, which corresponds to the DNA sequence alignment: *aln* $= \begin{pmatrix}\texttt{GACAGAGCAGTATTAGCTTCACAC}\\ \texttt{GAAAAC---GTATTA---TCAAAG}\end{pmatrix}$. That is, to preserve the reading frame, a robust alignment is accomplished translating the codon sequence into aminoacid sequence alignment.

Sequences A and $B' =$ GAAAAC---GTATTA---TCAAAG can also be represented as elements from group:

$$G_{A'} = C_g^{\mathrm{ACGT}} \oplus C_g^{\mathrm{TGCA}} \oplus C_{g+}^{\mathrm{ACGT}} \oplus (C_g^{\mathrm{ACGT}})^2 \oplus C_{g+}^{\mathrm{ACGT}} \oplus (C_g^{\mathrm{ACGT}})^2$$

This group is an element of the **Ab** category $\mathcal{C}_{A'}$, which is a subcategory of the $\boldsymbol{R_{A'}}$**-Mod** category over the ring $R_{A'} = \left(\mathbb{Z}_{2^6}\right)^2 \otimes \mathbb{Z}_{5^3} \otimes \left(\mathbb{Z}_{2^6}\right)^2 \otimes \mathbb{Z}_{5^3} \otimes \left(\mathbb{Z}_{2^6}\right)^2$. The group isomorphism $F_B : G_B \to G_{B'}$ is the functor that maps DNA sequences from group $G_B \in \mathcal{C}_N$ into an element from group $G_{B'} \in \mathcal{C}_R$ (see Appendix B). That is, for all element $b = (X_1, X_2, X_3, X_4, X_5, X_6)$ ($b \in G_B$) there is a unique element $b' = (X_1', X_2', 0, X_3', X_4', 0, X_5', X_6')$ ($X_i' = X_i$ and $b' \in G_{B'}$).

Also, there is an injective morphism $F_A : G_A \to G_{A'}$ that transforms each element $a = (X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$ ($a \in G_A$) into a unique element $a' = (X_1', X_2', X_3', X_4', X_5', X_6', X_7', X_8')$ ($a' \in G_{A'}$), which is evident since $C_g \subset C_{g+}$ and, consequently, codons are preserved, i.e., $X_i' = X_i$ and G_A is isomorphic to the image $F_A\left(G_A\right)$. The homomorphism $F_{A'} : G_A \to G_{A'}$ is also a functor which maps elements from the $\boldsymbol{R_A}$**-Mod** category over the ring $R_A = \otimes_8 \mathbb{Z}_{2^6}$ into the $\boldsymbol{R_{A'}}$**-Mod** category. Notice that $F_B\left(B\right)$ is a subgroup of $F_A\left(A\right)$.

In practice, for the sake of computational genomics implementations, the aligned DNA sequences A and B can be represented by the numerical vectors a $a = (9,32,24,56,60,27,28,5)$ and $b = (8,1,56,60,28,1)$, respectively, with coordinates on $\mathbb{Z}_{2^6}$. The application of the morphisms F_A and F_B permits the new representations: $a' = \big((9,32) \in \mathbb{Z}_{2^6}, 66 \in \mathbb{Z}_{5^3}, (56,60) \in \mathbb{Z}_{2^6}, 69 \in \mathbb{Z}_{5^3}, (28,5) \in \mathbb{Z}_{2^6}\big)$ and $b' = \big((8,1) \in \mathbb{Z}_{2^6}, 0 \in \mathbb{Z}_{5^3}, (56,60) \in \mathbb{Z}_{2^6}, 0 \in \mathbb{Z}_{5^3}, (28,1) \in \mathbb{Z}_{2^6}\big)$, respectively. The group homomorphism φ with matrix representation with diagonal elements $\big((8,2) \in \mathbb{Z}_{2^6}, 0 \in \mathbb{Z}_{5^3}, (1,1) \in \mathbb{Z}_{2^6}, 0 \in \mathbb{Z}_{5^3}, (1,1) \in \mathbb{Z}_{2^6}\big)$ maps sequence a' into b', *i.e.*, $\varphi(a') = b'$:

$$(9,32,66,56,60,69,28,5)\begin{bmatrix} 8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 13 \end{bmatrix} = \begin{pmatrix} 8 \\ 1 \\ 0 \\ 56 \\ 60 \\ 0 \\ 28 \\ 1 \end{pmatrix}$$

Where the third and sixth rows are computed modulo 125 and the rest modulo 64. The group homomorphism $h: G_{B'} \to G_A$ that accomplish the mapping $h(b') = a'$ is computed as:

$$(8,1,0,56,60,0,28,1)\begin{bmatrix} 58 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 32 & 66 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 57 & 0 & 0 & 0 & 0 & 69 & 0 & 5 \end{bmatrix} = \begin{pmatrix} 9 \\ 32 \\ 66 \\ 56 \\ 60 \\ 69 \\ 28 \\ 5 \end{pmatrix}$$

Or by means of the affine transformation:

$$\left((8,1,0,56,60,0,28,1)\begin{bmatrix} 58 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 32 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 5 \end{bmatrix} \bmod 64 + \begin{pmatrix} 118 \\ 0 \\ 66 \\ 0 \\ 0 \\ 69 \\ 0 \\ 0 \end{pmatrix} \right) \bmod 125 = \begin{pmatrix} 9 \\ 32 \\ 66 \\ 56 \\ 60 \\ 69 \\ 28 \\ 5 \end{pmatrix}$$

In summary, a future theoretical development in the framework of category theory opens new horizons for the analysis of the mutational process in a wider computational genomic scenario not previously studies in molecular evolutionary biology.

4 Discussions

The encoding of the physicochemical relationships between nucleotides (nitrogenous bases) in the DNA double helix in terms of group operations permits a mathematical representation of genome architecture interpretable in a molecular evolutionary context. The group representations of the genetic code are logically extended from protein-coding DNA regions to the entire genome. As shown in Fig 1, the Abelian group representation of genomic regions into the direct sum of Abelian p-group is only one of several steps addressed to get better understanding on how genomes are built.

The advantage on using group representations is that there exists a well-established mathematical development that leads to an objective study of the genome architecture in a molecular evolutionary context, through the analysis of mutational events in terms of group homomorphisms: endomorphisms, automorphisms, and translations. On this scenario the analysis of group homomorphisms permits us the uncovering of stochastic rules constraining the local architecture on genes and genomic regions. The goal is unveiling hidden genomic architecture and rules hard to be detected by current experimental approaches. All the information required can be retrieved from the MSA of DNA sequences, which is particularly relevant for poorly annotated genomes.

Examples shown in Figs 3 to 4 indicates that 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 homocyclic or cyclic groups representing a 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 $\left(C_g,+\right)\cong\left(\mathbb{Z}_{64},+\right)$, in protein coding regions, and by any combination of groups $(\mathfrak{B},+) \cong (\mathbb{Z}_2^2,+)$, $(\mathfrak{B}^2,+) \cong (\mathbb{Z}_2^4,+)$ or some quotient group like $C_G / G_{\mathrm{GGA}} \cong (\mathbb{Z}_2,+)$ in non-protein coding regions.

Notably, the genetic code Abelian group $(C_{G+},+) \cong (\mathbb{Z}_5^3,+)$ is enough for an algebraic representation of the genome population from the same species or close related species. However, such a decomposition leads to a poor description of local architecture that, as suggested in Figs. 3 to 6, can mask relevant biological features. Figure 3 to 6 illustrate the basic Abelian group representations for further analysis of genome architecture through the study of the mutational events, as essential transformations inherent to the molecular evolutionary process, in terms of endomorphisms and automorphisms, elements of the endomorphism ring.

For the sake of reader's comprehension, the examples on the group representation of genomic regions presented here are simple. However, the analysis demands for the development of novel computational algebraic approaches to study the genomic architecture. Unlike to traditional computational algebra, we can take advantage of the group isomorphisms, which permits decreasing the computational complexity by avoiding symbolic computation. Nevertheless, results presented here show that the architecture of genome region in an entire population can be quantitively studied in the framework of Abelian group theory.

From several examples provided here, it is clear that there exists a language for the genome architecture unveiled when represented it in terms of sums of finite Abelian groups, which can be further studied with the application of methods from category theory, the potential success of which has been proven in programming languages and in linguistic [52]. The future developments of genome annotation from several species can certainly lead to the discovery of logical rules from such a language, finding the viable variations in the populations. The identification of quotient groups (at larger scale) can permit the stratification of large genome population into equivalence classes (quotient subgroups) corresponding to individual subpopulations, each one of them carrying particular viable variations of species genome architecture.

As indicated in reference [18], 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 [5,53]. 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 from the model where base *D* stands for the methylated cytosine and adenine. As suggested in Fig 7**a** and **b**, a precise decomposition of methylation motif into the direct sum of Abelian finite groups can lead to their classification into unambiguous equivalence classes. The group structure of the methylation regulatory regions: **GATC**TTTTATGC and GGTTAAAA**GATC,** both represented by the homocyclic group on $\left(\mathbb{Z}_5^3\right)^4$, breaks from the monotone homocyclic group representation of the region in terms of cyclic groups on $\mathbb{Z}_4$ (Fig 7**a** and **b**). The group representation of protein-coding regions (or base-triplet sequences) as numerical vectors with coordinates on $\mathbb{Z}_{5^3}$ (Fig 7**c** and **d**) facilitates the analysis of methylation changes represented as group endomorphism/automorphisms of the cyclic group on $\mathbb{Z}_{5^3}$.

Results indicate that, as a consequence of the genetic code constraints and the evolutionary pressure on protein-coding regions, stochastic-deterministic logical rules can be inferred on a large enough sample-size from a gene/genomic-region population. Such a stochastic-deterministic rules lead to specific applications of Theorem 1 and Eq.8, consequently, the analysis of mutational process on each group, subgroup, and coset. For example, mutational events on a MSA column (identified) from class YHH (with discriminatory classification power as shown in Fig 8) where the second and third DNA bases remain invariant (H) and the first base are pyrimidines (Y) experiencing transition mutations (across individuals sequences) are represented by automorphisms on a subgroup (from the genetic code Abelian subgroup $\left(C_G,+\right)$) defined on the set $\{\text{THH},\text{CHH}\}$isomorphic to $\left(\mathbb{Z}_2,+\right)$ [20].

Figure 9 provides illustrative example that motivates further applications of based machine-learning bioinformatic approaches to unveil the subjacent logic to the genome architecture and its association with the DNA cytosine/adenine methylation patterning found on individual populations and the changes (repatterning) induced by, e.g., environmental changes, aging process and diseases, which is of particular interest in genomic medicine [54]. Machine learning applications on MSA involving large sample size of genomic regions from populations of different species can unveil further decompositions into the direct sum of Abelian groups, which do not depend on our current knowledge of the annotated genomes. As suggested in Fig 9, we can expect that most of the hiding genomic DNA sequence motif can be unveiled by studying the molecular evolutionary (mutational) process in a genome population through the lens of the endomorphism ring. In other words, as a consequence of the injective relationship: *DNA sequence* → *3D chromatin architecture* [3,4,6], fixed mutational events (in organismal populations) on DNA sequence motifs involved in the 3D chromatin architecture are under evolutionary pressure, biophysically and biochemically constrained to preserve the chromatin biological functions.

Results shown in Figs. 8 and 9 also suggest deep implications of Baer-Kaplansky theorem on the genome architecture unknown by the current knowledge and understanding of genome annotation, which currently relies on the DNA sequence itself. Concretely, on an evolutionary context, the fact that two genomic regions from two different species are almost identical and, event would encode for the same functional protein, does not necessarily imply that they hold to the same genome architecture. The evolutionary pressure in both such hypothetical regions must be same, which implies that the regions experience the same type of mutational events in terms of automorphism/endomorphism representations.

For example, let’s suppose that the results shown in Fig 9 were derived from a large sample size (large enough to derive statistically significant rules), then the rule “A1 $\wedge$ R3 $\wedge$ (YHH $\vee$ HHY) → human” (Fig 9) implies that the gene regions of BRCA1 from human and non-human primates do not belong to the same equivalent class of genomic region. In particular, since the endomorphism rings $\mathfrak{R}\left(G_{human}^{\mathrm{BRCA1}}\right)$ and $\mathfrak{R}\left(G_{non-human}^{\mathrm{BRCA1}}\right)$ on the Abelian groups $G_{human}^{\mathrm{BRCA1}}$ and $G_{non-human}^{\mathrm{BRCA1}}$ defined on the human and non-human primates BRCA1 genes, respectively, are not isomorphic, then according to the Baer-Kaplansky theorem groups G_{human}^{BRCA1} and $G_{non-human}^{\text{BRCA1}}$ are also not isomorphic. Hence, region architectures of BRCA1 gene in human and non-human primates are (in this hypothetical scenario) implicitly different, which is not obvious to human eyes from their MSA (see supporting information).

Results presented here would have considerable positive impact on current molecular evolutionary biology, which heavily relies on subjective evolutionary null hypotheses about the past. As suggested in reference [11], the genomic Abelian groups open new horizons for the study of the molecular evolutionary stochastic processes (at genomic scale) with relevant biomedical applications, founded on a deterministic ground, which only depends on the physicochemical properties of DNA bases and aminoacids. In this scenario, the only molecular evolutionary hypothesis needed about the past is a fact, the existence of the genetic code.

Remarkably, further studies applying the theory presented here do not require for special experimental datasets but for the DNA sequences of the genomic regions under scrutiny. Although the accuracy of the predictions depends on the sample size, the number of sequenced genomes stored in the databases grows year-after-year. Large samples of DNA sequences (from homolog genomic regions) from at least two or more species facilitate application of Baer-Kaplansky theorem and further studies applying methods of Categorical theory to unveil the grammar embedded in the DNA sequences.

The theory and concretes examples provided here make explicit the basic foundation for a further unprecedented application of the last advances in Abelian group theory incorporating methods from Category theory, where groups and group homomorphisms (in our context: mutational events) are the main players, which have the potential to discover unsuspected features of the genome architecture, opening new horizon to the genomic taxonomy of species in accordance with the state-of-the-art in mathematics, logic, and computational sciences. In other words, these applications have the potential to elevate the genomic studies from the current descriptive level to the vanguard level marked in the frontier of science by mathematics, physics, and computational sciences.

5 Conclusions

Results to date indicate that the genetic code and 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 starting from genomic regions till cover whole chromosomes. 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 [30,55].

Results suggest that the architecture of current population genomes is quite far from randomness and obeys stochastic-deterministic rules. The nexus between the Abelian finite group decomposition into homocycle Abelian *p*-groups and the endomorphism ring paved the ways to unveil unsuspected stochastic-deterministic logical propositions ruling the ensemble of genomic regions and sets the basis for a novel algebraic taxonomy of the species, which is not limited by our current biological knowledge.

In the context of evolutionary comparative genomics, the theory presented here open new horizons for the application of Group theory including methods of Category theory, which have the potential to unveil hidden features and rules inherent to the genome architecture, leading to an unprecedented understanding on how genomes are built.

We believe that the mathematical formalism proposed here sets the theoretical ground for a further development in genomics, transitioning the field from a fully empirical science to a predictive science, where the theoretical and empirical research coexist in a tight positive feedback loop, a development stage only reached so far in the field of physics.

All the above claims are feasible, only limited by our computational power and the availability of samples of sequenced genomes from the same species and from multiple species.

At this point we emphasize that *an accurate understanding of the genome architecture and population's structure, on a formal mathematical framework, is as essential for the future of genetic* *engineering and genome editing as the physics of architecture is to the design of sturdy and stable energy-efficient building.*

6 Appendix A. Genetic code algebraic structures defined on the base and codon sets

An Abelian group structure $(B,+)$ is a set $\boldsymbol{B}$ together with a binary operation '+' that combines any two elements $a \in B$ and $b \in B$ to form another element of $c \in B$, denoted $a+b=c$, which satisfy the following axioms:

1) Associativity. For all $a,b,c \in B$, the equality $(a+b)+c=a+(b+c)$ holds.
2) Identity. There exists an element $e \in B$ named identity element of B, such that for any $a \in B$, the equality $a+e=a$ holds.
3) Commutativity**.** For all $a,b \in B$, the equality $a+b=b+a$

The Abelian groups considered here are finite cyclic groups $(G,+)$ isomorphic to the Abelian group defined on the set of integers modulo *n*, denoted as $\mathbb{Z}_n(\mathbb{Z}/n\mathbb{Z})$. That is, the integers $1,2,3,\ldots,n-1$ form a cyclic group of order *n* under addition (modulo *n*) and 0 as the identity element. This group will be denoted as $(\mathbb{Z}_n,+)$. However, for the sake of simplicity in the figures it will be denoted simply as $\mathbb{Z}_n$, i.e., without making distinction between the set $\mathbb{Z}_n$ and group structure defined on it. The particular interest for the current work is the Abelian *p*-group derived when $n = p^{\alpha}$ where *p* is a prime number and α an integer. The group operations defined on the set of bases or on the codon set are associated to physicochemical properties of DNA bases (see the next sections).

Homomorphisms and isomorphisms

In modern algebra, a group homomorphism is a map $f: A \to B$ between two group structures $(A,\bullet)$ and $(B,\circ)$ such that for all $a,b \in A$ holds: $f(\alpha_1 \bullet \alpha_2) = f(\alpha_1) \circ f(\alpha_2) = \beta_1 \circ \beta_2$, where

$\beta_1, \beta_2 \in B$. A group isomorphism is a one-to-one correspondence (mapping) between two sets that preserves binary relationships between elements of the sets. That is, an isomorphism is a homomorphism holding the inverse mapping: $f^{-1}(\beta_1 \circ \beta_2) = f^{-1}(\beta_1) \bullet f^{-1}(\beta_2) = \alpha_1 \bullet \alpha_2$. For example, since there exists only one cyclic group with four elements up to isomorphism, for each one of the 24 cyclic group $(\mathfrak{B}, +_b)$ defined on the set of bases $\mathfrak{B} = \{\mathrm{A,C,G,T}\}$ ([17,18]) there exists a one-to-one mapping f such that for each base $\beta \in \mathfrak{B}$ there is an integer $\iota \in \mathbb{Z}_4$ such that $f(\beta) = \iota$ and:

1. $f(\beta_1 +_b \beta_2) = f(\beta_1) + f(\beta_2) = \iota_1 + \iota_2$, $\beta_1, \beta_2 \in \mathfrak{B}$ and $\iota_1, \iota_2 \in \mathbb{Z}_4$.
2. The inverse mapping $f^{-1}(\iota_1 + \iota_2) = f^{-1}(\iota_1) +_b f^{-1}(\iota_2) = \beta_1 +_b \beta_2$

To highlight the fact that the sum operations are defined on different ways on the sets $\mathfrak{B}$ and $\mathbb{Z}_4$, we have used the symbols '$+_b$' and '+', respectively. However, for sake of brevity of the symbolic notation, such knowledge will be considered implicit, writing simply '+'. Then, we said that groups $(\mathfrak{B}, +_b)$ and $(\mathbb{Z}_4, +)$ are isomorphic; in symbols $(\mathfrak{B}, +_b) \cong (\mathbb{Z}_4, +)$. f and its inverse f^{-1} are called isomorphisms. If f (and its inverse f^{-1}) is a mapping from a group into itself, then f is called an automorphism. A mapping g, not necessarily one-to-one, of the elements from a group into itself is called a group endomorphism. An endomorphism that is also an isomorphism is an automorphism.

A ring algebraic structure is obtained when together with the sum operation "+" (as defined above) a new operation "$\cdot$" is defined on the set B holding the properties:

1. Associativity: $(a \cdot b) \cdot c = a \cdot (b \cdot c)$ for all $a, b, c \in B$
2. Multiplication is distributive with respect to addition:
   a. $(a + b) \cdot c = (a \cdot c) + (b \cdot c)$ for all $a, b, c \in B$ (right distributivity).
   b. $c \cdot (a + b) = c \cdot a + c \cdot b$ for all $a, b, c \in B$ (left distributivity).

As it is shown in the next section, these algebraic structures have been defined on the genetic code. In particular, the ring $\left(\mathbb{Z}_{2^6}, +.\cdot\right)$ and endomorphism ring (section 3.1) has been defined and studied on the genetic code [18].

Appendix B. Category

Category theory is a general mathematical theory of structures and of systems of structures that occupy a central position in contemporary mathematics, theoretical computer science, and linguistics [56].

Definition: A category $\mathcal{C}$ can be described as a collection of objects $\mathcal{O}$ satisfying the following three conditions:

1) *Morphism*: For every pair X, Y of objects, there is a set $Hom(X, Y)$ called the *morphisms* from X to Y in $\mathcal{C}$. If f is a morphism from, we write $f: X \to Y$.
2) *Identity*: For every object X, there exists a morphism id_X in $Hom(X, Y)$, called the *identity* on X (also denoted as 1_X).
3) *Composition*: For every triple X, Y, and Z of objects, there exists a partial binary operation from $Hom(X, Y) \times Hom(Y, Z)$ to $Hom(X, Z)$, called the composition of morphisms in $\mathcal{C}$. If $f: X \to Y$ and $g: Y \to Z$, this composition is written as the mapping $(g \circ f): X \to Z$.

Identity, morphisms, and composition satisfy two axioms:

Associativity: If $f: X \to Y$, $g: Y \to Z$, and $h: Z \to W$, then $h \circ (g \circ f) = (h \circ g) \circ f$.

Identity: If $f: X \to Y$, then $f_X \circ f = f$ and $f \circ f_X = f$.

Definition: A functor F is a function between two categories $\mathcal{C}$ and D which maps objects to objects and morphisms to morphisms. That is:

- For each $X \in \mathcal{C}$ there is an object $F(Y) \in D$

- For each morphism $f: X \to Y$ in $\mathcal{C}$ there is morphism $F(f): F(X) \to F(Y)$ in D such that the following conditions hold:
  i. $F(g \circ f) = F(g) \circ F(f)$ for all morphisms $f: X \to Y$ and g: $X \to Y$ in $\mathcal{C}$
  ii. $F(id_X) = id_{F(X)}$ for all $X \in \mathcal{C}$.

7 Supporting Information

A summary with the reported genetic code Abelian groups relevant for the current study is provided as supporting information in a file named: *Supporting_Information.docx*.

All the data, computational and statistical analyses can be reproduced following the R scripts provided in tutorials available at the GenomAutomorphism R package website https://genomaths.github.io/genomautomorphism/. In particular, data and R scripts used in the computation of automorphisms and the decision tree from Fig 9 are available within GenomAutomorphism R package and in a tutorial at:

https://genomaths.github.io/genomautomorphism/articles/automorphism_and_decision_tree.html.

8 References

[1] E. V. Koonin, Evolution of genome architecture, Int. J. Biochem. Cell Biol. 41 (2009) 298–306. https://doi.org/10.1016/j.biocel.2008.09.015.

[2] S. Whalen, R.M. Truty, K.S. Pollard, Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin, Nat. Genet. 48 (2016) 488–496. https://doi.org/10.1038/ng.3539.

[3] J. Nuebler, G. Fudenberg, M. Imakaev, N. Abdennur, L.A. Mirny, Chromatin organization by an interplay of loop extrusion and compartmental segregation, Proc. Natl. Acad. Sci. U. S. A. 115 (2018) E6697–E6706. https://doi.org/10.1073/pnas.1717730115.

[4] M.J. Rowley, V.G. Corces, Organizational principles of 3D genome architecture, Nat. Rev. Genet. 19 (2018) 789–800. https://doi.org/10.1038/s41576-018-0060-8.

[5] A. Piazza, W.D. Heyer, Homologous Recombination and the Formation of Complex Genomic Rearrangements, Trends Cell Biol. 29 (2019) 135–149. https://doi.org/10.1016/j.tcb.2018.10.006.

[6] H. Zheng, W. Xie, The role of 3D genome organization in development and cell differentiation, Nat. Rev. Mol. Cell Biol. 20 (2019) 535–550. https://doi.org/10.1038/s41580-

019-0132-4.

[7] T.D. Schneider, Evolution of biological information, Nucleic Acids Res. 28 (2000) 2794–9. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=102656&tool=pmcentrez&rendertype=abstract.

[8] H.P. Yockey, Origin of life on earth and Shannon's theory of communication, Comput. Chem. 24 (2000) 105–123. https://doi.org/10.1016/S0097-8485(99)00050-9.

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

[10] R. Sanchez, S.A. Mackenzie, Information thermodynamics of cytosine DNA methylation, PLoS One. 11 (2016) e0150427. https://doi.org/10.1371/journal.pone.0150427.

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

[12] R. Sánchez, E. Morgado, R. Grau, A genetic code Boolean structure. I. The meaning of Boolean deductions, Bull. Math. Biol. 67 (2005) 1–14. https://doi.org/10.1016/j.bulm.2004.05.005.

[13] H.P. Yockey, Information theory, evolution and the origin of life, Inf. Sci. (Ny). 141 (2002) 219–225. https://doi.org/10.1016/S0020-0255(02)00173-1.

[14] R. Sanchez, R. Grau, A genetic code Boolean structure. II. The genetic information system as a Boolean information system, Bull. Math. Biol. 67 (2005) 1017–1029. https://doi.org/10.1016/j.bulm.2004.12.004.

[15] D.. Andrews, M.L. Boss, DNA code may transmit "Superinformation," Yale Sci. Mag. 45 (1971) 50–51.

[16] G.S. Chirikjian, Algebraic and Geometric Coding Theory, in: Stoch. Model. Inf. Theory, Lie Groups, Vol. 2 Anal. Methods Mod. Appl., Birkhäuser Boston, Boston, 2012: pp. 313–336. https://doi.org/10.1007/978-0-8176-4944-9_9.

[17] M. V José, E.R. Morgado, R. Sánchez, T. Govezensky, The 24 Possible Algebraic Representations of the Standard Genetic Code in Six or in Three Dimensions, Adv. Stud. Biol. 4 (2012) 119–152. http://www.m-hikari.com/asb/asb2012/asb1-4-2012/joseASB1-4-2012-1.pdf.

[18] R. Sanchez, E. Morgado, R. Grau, Gene algebra from a genetic code algebraic structure, J. Math. Biol. 51 (2005) 431–457. https://doi.org/10.1007/s00285-005-0332-8.

[19] R. Sanchez, R. Grau, E. Morgado, A novel Lie algebra of the genetic code over the Galois field of four DNA bases, Math. Biosci. 202 (2006) 156–174.

https://doi.org/10.1016/j.mbs.2006.03.017.

[20] R. Sánchez, R. Grau, An algebraic hypothesis about the primeval genetic code architecture, Math. Biosci. 221 (2009) 60–76. https://doi.org/S0025-5564(09)00114-X [pii] 10.1016/j.mbs.2009.07.001.

[21] D.. Andrew, M.L. Boss, DNA code may transmit “Superinformation,” Chem. Eng. News. 48 (1970) 50–51.

[22] H.J. Danckwerts, D. Neubert, Symmetries of genetic code-doublets, J Mol Evol . 5 (1975) 327–332. https://doi.org/10.1007/BF01732219.

[23] M.O. Bertman, J.R. Jungck, Group graph of the genetic code, J. Hered. 70 (1979) 379–384. https://doi.org/10.1093/oxfordjournals.jhered.a109281.

[24] J.E.M. Hornos, Y.M.M. Hornos, Algebraic model for the evolution of the genetic code, Phys. Rev. Lett. 71 (1993) 4401–4404. https://doi.org/10.1103/PhysRevLett.71.4401.

[25] J.E.M. Hornos, Y.M.M. Hornos, M. Forger, Symmetry and Symmetry Breaking: an Algebraic Approach To the Genetic Code, Int. J. Mod. Phys. B. 13 (1999) 2795–2885. https://doi.org/10.1142/S021797929900268X.

[26] M. Forger, S. Sachse, Lie superalgebras and the multiplet structure of the genetic code. I. Codon representations, J. Math. Phys. 41 (2000) 5407–5422. https://doi.org/10.1063/1.533417.

[27] K. Tamura, M. Nei, Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees, Mol. Biol. Evol. 10 (1993) 512–26. http://www.ncbi.nlm.nih.gov/pubmed/8336541.

[28] R. Sanchez, GenomAutomorphism: Compute the automorphisms between DNA’s Abelian group representations. R package version 1.0.0, (2020). https://doi.org/10.18129/B9.bioc.GenomAutomorphism.

[29] L. Fuchs, Abelian groups, Publishing House of the Hungarian Academy of Sciences, 1958.

[30] L. Fuchs, Infinite Abelian Groups - Volume I, 1st Editio, Academic Press, 1970. https://doi.org/10.1007/978-3-319-19422-6.

[31] M. Yang, M.K. Derbyshire, R.A. Yamashita, A. Marchler-Bauer, NCBI’s Conserved Domain Database and Tools for Protein Domain Analysis, Curr. Protoc. Bioinforma. 69 (2020) 1–25. https://doi.org/10.1002/cpbi.90.

[32] A.B. Stergachis, E. Haugen, A. Shafer, W. Fu, B. Vernot, A. Reynolds, A. Raubitschek, S. Ziegler, E.M. LeProust, J.M. Akey, J.A. Stamatoyannopoulos, Exonic transcription factor binding directs codon choice and affects protein evolution., Science (80-.). 342 (2013) 1367–72. https://doi.org/10.1126/science.1243490.

[33] I. Reyna-Llorens, S.J. Burgess, G. Reeves, P. Singh, S.R. Stevenson, B.P. Williams, S. Stanley, J.M. Hibberd, Ancient duons may underpin spatial patterning of gene expression in

C 4 leaves, Proc. Natl. Acad. Sci. U. S. A. 115 (2018) 1931–1936. https://doi.org/10.1073/pnas.1720576115.

[34] V.K. Yadav, K.S. Smith, C. Flinders, S.M. Mumenthaler, S. De, Significance of duon mutations in cancer genomes, Sci. Rep. 6 (2016) 1–9. https://doi.org/10.1038/srep27437.

[35] R. Sanchez, R. Grau, E.R. Morgado, R. Sánchez, L.A. Perfetti, R. Grau, E.R.M. Morales, A New DNA Sequence Vector Space on a Genetic Code Galois Field, MATCH Commun. Math. Comput. Chem. 54 (2005) 3–28. http://match.pmf.kg.ac.rs/electronic_versions/Match54/n1/match54n1_3-28.pdf.

[36] Y. Lv, J. Gu, H. Qiu, H. Li, Z. Zhang, S. Yin, Y. Mao, L. Kong, B. Liang, H. Jiang, C. Liu, Whole-exome sequencing identifies a donor splice-site variant in SMPX that causes rare X-linked congenital deafness, Mol. Genet. Genomic Med. 7 (2019) e967. https://doi.org/10.1002/mgg3.967.

[37] A. Panigrahi, B.W. O'Malley, Mechanisms of enhancer action: the known and the unknown, Genome Biol. 22 (2021) 1–30. https://doi.org/10.1186/s13059-021-02322-1.

[38] X. Sheng, J. Wang, Y. Guo, J. Zhang, J. Luo, DNA N6-Methyladenine (6mA) Modification Regulates Drug Resistance in Triple Negative Breast Cancer, Front. Oncol. 10 (2021) 1–6. https://doi.org/10.3389/fonc.2020.616098.

[39] Q. Lin, J. wei Chen, H. Yin, M. an Li, C. ren Zhou, T. fang Hao, T. Pan, C. Wu, Z. ran Li, D. Zhu, H. fan Wang, M. sheng Huang, DNA N6-methyladenine involvement and regulation of hepatocellular carcinoma development, Genomics. 114 (2022) 110265. https://doi.org/10.1016/j.ygeno.2022.01.002.

[40] M. Zamora, C.A. Ziegler, P.L. Freddolino, A.J. Wolfe, A Thermosensitive, Phase-Variable Epigenetic Switch: pap Revisited, Microbiol. Mol. Biol. Rev. 84 (2020). https://doi.org/10.1128/mmbr.00030-17.

[41] R. Sanchez, S.A. Mackenzie, Integrative Network Analysis of Differentially Methylated and Expressed Genes for Biomarker Identification in Leukemia, Sci. Rep. 10 (2020) 2123. https://doi.org/10.1038/s41598-020-58123-2.

[42] G. Ivanov, Generalizing the Baer-Kaplansky Theorem, J. Pure Appl. Algebr. 133 (1998) 107–115. https://doi.org/10.1016/S0022-4049(97)00187-4.

[43] P.W. Gunning, E.C. Hardeman, Isoforms: Fundamental differences, Elife. 7 (2018) 1–3. https://doi.org/10.7554/eLife.34477.

[44] L.P. Iñiguez, G. Hernández, The evolutionary relationship between alternative splicing and gene duplication, Front. Genet. 8 (2017) 1–7. https://doi.org/10.3389/fgene.2017.00014.

[45] F.H.C. Crick, The Origin of the Genetic Code, J. Mol. Biol. 38 (1968) 367–379.

[46] A. Comnish-Bowden, Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984, Nucleic Acids Res. 13 (1985) 3021–3030.

[47] G.V.Kass, An Exploratory Technique for Investigating Large Quantities of Categorical Data, J. Roral Stat. Soc. 29 (1980) 119–127.

[48] S. Miyazawa, R.L. Jernigan, Residue-Residue Potentials with a Favorable Contact Pair Term and an Unfavorable High Packing Density Term, for Simulation and Threading, J Mol Biol. 256 (1996) 623–644. https://doi.org/10.1006/jmbi.1996.0114.

[49] R. Sanchez, Evolutionary Analysis of DNA-Protein-Coding Regions Based on a Genetic Code Cube Metric, Curr. Top. Med. Chem. 14 (2014) 407–417. https://doi.org/10.2174/1568026613666131204110022.

[50] P. Freyd, Abelian Categories: An Introduction to the Theory of Functors, Harper & Row, New York, 1964.

[51] B. Mitchell, The Full Imbedding Theorem, Amer. J. Math. 86 (1964) 619. https://doi.org/10.2307/2373027.

[52] Y. Maruyama, Category theory and foundations of life science: A structuralist perspective on cognition, BioSystems. 203 (2021) 1–20. https://doi.org/10.1016/j.biosystems.2021.104376.

[53] M. Yu, B. Ren, The three-dimensional organization of mammalian genomes, Annu. Rev. Cell Dev. Biol. 33 (2017) 265–289. https://doi.org/10.1146/annurev-cellbio-100616-060531.

[54] Y. Salameh, Y. Bejaoui, N. El Hajj, DNA Methylation Biomarkers in Aging and Age-Related Diseases, Front. Genet. 11 (2020) 1–11. https://doi.org/10.3389/fgene.2020.00171.

[55] L. Fuchs, Infinite Abelian Groups, Volume 2, Academic Press, 1973.

[56] J.-P. Marquis, Category Theory, in: E.N. Zalta (Ed.), {Stanford} Encycl. Philos., {F}all 202, Metaphysics Research Lab, Stanford University, 2021.