THE EMERGENT GENOME

Edward A. Ruiz-Narváez[1]

[1] Department of Nutritional Sciences, University of Michigan School of Public Health, Ann Arbor, MI

Declarations of interest: none

Corresponding author: Edward A. Ruiz-Narváez, Department of Nutritional Sciences, University of Michigan School of Public Health, 1860 SPH I, 1415 Washington Heights, Ann Arbor, MI 48109. Phone: 734-647-0623. Email: eruiznar@umich.edu

ABSTRACT

Current paradigm equates an organism's genome with its complete DNA sequence. However, results from omics research show that the genome is more than the DNA sequence. For example, sequence alone does not determine multi-functionality of regulatory elements (e.g. enhancers, insulators). In addition, identity of genomic elements depends on cellular and temporal context. Based on these findings, the present work advances the hypothesis that the genome is an emergent entity resulting from epigenomics mechanisms. The genome can be understood as the mapping of identity-functions to elements along the DNA molecule. The mapping can vary across cellular types and developmental times. As consequence, the same organism can have multiple genomes regardless of the underlying DNA sequence. The proposed theory has major implications for the study of the hereditary basis of phenotypic traits, including diseases, and offers a new framework for future research.

“From now onwards space by itself and time by itself will recede completely to become mere shadows and only a type of union of the two will still stand independently on its own.”

Hermann Minkowski, lecture at the 80th Meeting of German Natural Scientists, Cologne, September 21, 1908

INTRODUCTION

Current theory and practice in genetics treat the genome and epigenome as two related although different entities. For example, the epigenome is usually defined as “… a multitude of chemical compounds that can tell the genome what to do” (National Human Genome Research Institute, https://www.genome.gov/27532724/epigenomicss-fact-sheet/), the genome being “… an organism’s complete set of DNA, including all of its genes” (U.S. National Library of Medicine, https://ghr.nlm.nih.gov/primer/hgp/genome). Popular metaphors compare the epigenome to an electrical switch turning on and off a light bulb (see for example https://www.mpg.de/9910690/epigenetic-switch-obesity), or to annotations on a music score to change musical performance (see for example *Epigenome: The symphony in your cells*, http://www.nature.com/news/epigenome-the-symphony-in-your-cells-1.16955; and (Burris and Baccarelli, 2014)). Implicit on these definitions and metaphors is the ontological priority of the genome over the epigenome. Before any regulation could take place, there must be genes to be regulated. Before we turn on or turn off a light bulb, there must be a light bulb. Before we annotate a music score, there must be a music score. In other words, according to these definitions and metaphors, we could have a genome without an epigenome but not an epigenome without a genome. However, as I will discuss below, the current paradigm has both theoretical and practical limitations. I will argue that 1) neither the epigenome nor the genome can exist independently from each other, 2) the genome is an emergent entity resulting from epigenomics mechanisms, and 3) the epigenome and genome are part of a larger entity; the EpG^2 (EpiGenome-Genome) system.

THE GENOME DOES NOT EXIST WITHOUT THE EPIGENOME

In day-to-day talk is common use to identify the genome as the complete DNA sequence of an organism. For example, the international project that determined the complete sequence of the human DNA was known as the Human Genome Project (HGP), and terms such as "whole-genome sequencing" (WGS) are used to refer to experiments that sequence the totality of an organism's DNA. However, if we accept that an organism's genome is composed all of its genes, as well as regulatory (e.g. enhancers, insulators, promoters), structural (e.g. loop anchors, topologically associating domain (TAD) boundaries), and others still unknown functional elements then, as I will argue with some examples below, the genome is more than the DNA sequence. I will propose that the genome is an emergent entity, which results from epigenomics mechanisms.

What is a gene?

"What then is time? If no one asks me, I know what it is. If I wish to explain it to him who asks, I do not know." (Augustine, Confessions, Book XI, Chapter 14)

Modern biologists have a similar problem as Augustine's. We talk about genes in our daily professional practice. However, if we are pressed to define and explain what a gene is, we run into difficulties sooner rather than later. As discussed in details by others (Carlson, 1991; Gerstein et al., 2007; Griffiths and Stotz, 2006; Pesole, 2008; Portin and Wilkins, 2017; Scherrer and Jost, 2007), classical definitions of the "gene" concept fail to capture the complexity revealed by results from current research. Nested and overlapping genes, alternative and trans RNA splicing, extensive and continuous transcription along the DNA molecule are just a few examples of the phenomena that challenge our traditional understanding of the gene. I will not propose a new definition of the "gene" in this work, as many other authors have advanced novel ideas in how to redefine the "gene" in light of our current knowledge (Carlson, 1991; Gerstein et al., 2007; Griffiths and Stotz, 2006; Pesole, 2008; Portin and Wilkins, 2017; Scherrer and Jost, 2007). What these new definitions have in common is 1) the DNA sequence does not completely determine what a gene is or even where a particular gene is located in the genome, and 2) functional products (e.g. non-coding RNAs, proteins) are necessary for gene identity and location. In other words, the gene is in some sense a multidimensional and emergent entity involving material substrates (DNA, RNA, and protein) as well as the processes connecting them.

Promoter or enhancer? Insulator or promoter? Enhancer or insulator? What is it out there?

Identity (what it is) of regulatory and structural elements in the genome is tightly linked to their function (what it does). We say, for example, that promoters are proximal DNA sequences upstream of a gene that specify transcription start; enhancers are distal elements that increases transcription of genes; silencers suppress gene expression, and so forth (Maston et al., 2006). Recent results show that genomic elements are rather multifunctional: enhancers may act as silencers and vice versa (Kolovos et al., 2012), promoters and enhancers have exchangeable functions (Andersson, 2015; Kim and Shiekhattar, 2015), insulators may behave as promoters (Wei and Brennan, 2001), and tRNA genes may serve as insulators as well (Raab et al., 2012; Van Bortle and Corces, 2012). As I discuss below, these observations have unappreciated implications for our understanding of what is the genome.

Let us say there is a genomic element that have both enhancer and promoter activities, does this mean the element is 1) a promoter with added enhancer activity, 2) an enhancer with added promoter activity, or 3) both an enhancer and promoter with their corresponding activities. Implicit on these three alternatives is an absolute notion of genome, which stems from the traditional view of equating an individual's genome with its complete DNA sequence. According to this thinking, a person's genome is a fixed entity that, excepting somatic mutations and rearrangements, does not varies over time or cell type. This is the meaning of the usual expression saying that all cells of an individual have the same genome. However, if the genome is the set of all genes, regulatory and boundary elements, and any new functional elements yet to be discovered it is clear, from available evidence, that the genome is not the same as the complete DNA sequence. First, sequence alone does not determine identity or function of genomic elements. Instead, identify-function of genomic elements depends on the underlying DNA sequence, DNA-bound proteins (e.g. transcription factors, CTCF, etc.), cellular context, and developmental stage (Andersson, 2015; Erceg et al., 2017; Fourel et al., 2004; Palstra and Grosveld, 2012). Second, from an evolutionary point of view, DNA sequence conservation does not completely correlate with identity-function of genomic elements. In vertebrates, numerous enhancers have conserved function but divergent sequences (Yang et al., 2015), and new genomic elements may emerge without change of the underlying DNA sequence (i.e. exaptation of ancestral DNA) (Domene et al., 2013; Rebeiz and Tsiantis, 2017; Villar et al., 2015). At last, formation of new enhancers may result from the overexpression of transcription factors without changes in the DNA sequence (Hnisz et al., 2013; Shin, 2018). Although, this phenomena has been observed for now only in conditions such as cancer (Hnisz et al., 2013; Shin, 2018), it opens the possibility that in normal situations, genomic elements may emerge depending on concentrations and combinations of transcription factors, other DNA-bound proteins, and cellular context. In summary, 1) the genome is not a fixed entity corresponding to the complete DNA sequence of an organism, instead 2) the genome emerges as part of a new dynamic entity that we call the EpG^2 system.

THEORY OF THE EpG^2 SYSTEM

The following requires a basic knowledge of set and topology theory. However, main conclusions will be presented in an intuitive way. Consider the whole DNA molecule(s) of an individual composed of m genetic (i.e. sequence) elements. We will define:

Definition 1

- G is the set of n genetic elements in the whole DNA of an organism.
- G = $\{g_1, g_2, \ldots, g_n\}$, where g_i is the i-th genetic element

Definition 2

- (G, τ) is the genetic space with discrete topology, which means every possible subset of G – including the empty set ϕ and G itself are open sets– is open. For example, let us assume a genetic set G with only three elements: G = $\{g_1, g_2, g_3\}$. Then, the topological space (G, τ) will include all possible subsets of G: (G, τ) = $\{\phi, \{g_1\}, \{g_2\}, \{g_2\}, \{g_1, g_2\}, \{g_1, g_3\}, \{g_2, g_3\}, \{g_1, g_2, g_3\}\}$.
- We will use the notation $G_{ij\ldots k}$ to represent the $\{g_i, g_j, \ldots, g_k\}$ open set. According to this notation, the above topological space (G, τ) can be written as (G, τ) = $\{\phi, G_1, G_2, G_3, G_{12}, G_{13}, G_{23}, G_{123}\}$.

Definition 3

- F is the set of biologic identity-functions of an organism.
- F = $\{f_1, f_2, \ldots, f_m\}$, where f_i is the i-th biologic identity-function.

Definition 4

- $\pi\colon F \rightarrow (G, \tau)$ is an injective mapping from the F set to the topological space (G, τ) that is, if two identify-functions f_a and f_b map to the same open set $G_{ij\ldots k}$ then f_a and f_b are the same identity-function. In other words, for a given π mapping two different identity-functions cannot map to the same open set. Symbolically, $\forall f_a, f_b \in F, \pi(f_a) = \pi(f_b) \Rightarrow f_a = f_b$.
- For example, suppose the G set has three genetic elements, and the F set has three identity-functions. A possible π mapping from F to the topologic space (G, τ) is given by **Figure 1**

FIGURE 1

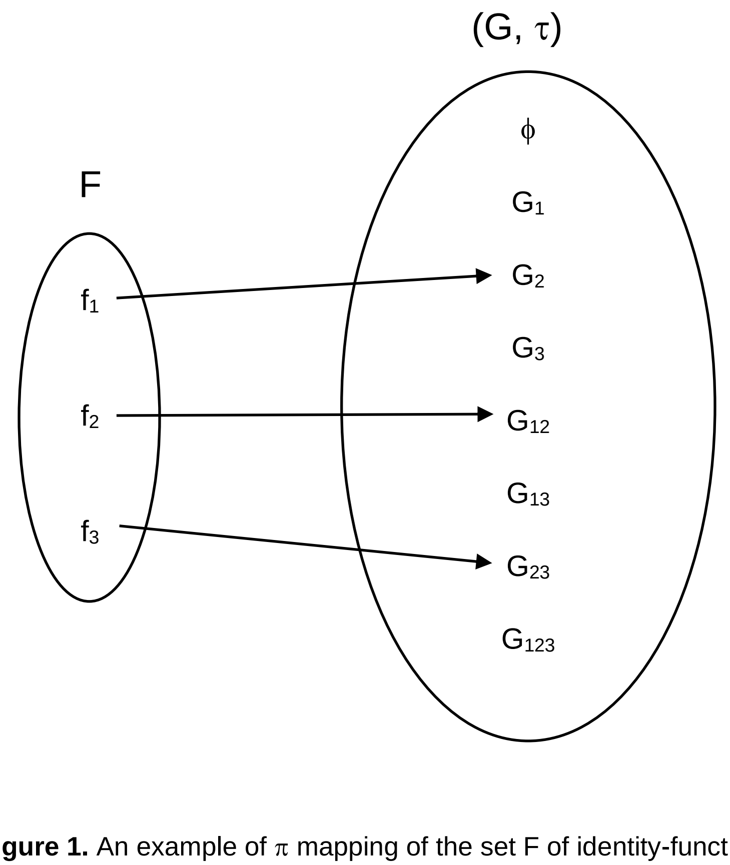


Figure 1. An example of π mapping of the set F of identity-functions to the genetic topological space (*G*, τ). In particular, $\pi(f_1) = G_2$, $\pi(f_2) = G_{12}$, and $\pi(f_3) = G_{23}$.

Matrix representation

The mapping shown in **Figure 1** can be represented as an m x n matrix **H**, where m (number of rows) equals the number of identity-functions, and n (number of columns) equals the number of genetic elements. The matrix element h_{ij} is equal to one if the j-th genetic element is included in the open set that is the image of the i-th identity-function, and equal to zero otherwise. Symbolically,

$h_{ij} = 1$ if $g_j \in \pi(f_i)$

$h_{ij} = 0$ if $g_j \notin \pi(f_i)$.

The above π mapping can be represented by the matrix

$$\boldsymbol{F} \overset{\boldsymbol{\pi}}{\to} (\boldsymbol{G}, \boldsymbol{\tau}) \equiv \ \boldsymbol{H} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}$$

where the i-th row vector $\boldsymbol{r_i}$ denotes the open set $\pi(f_i)$. From the above ***H*** matrix, the first identity-function (first row-vector) maps to the open set composed of the second genetic element, the second identity-function (second row-vector) maps to the open set composed of the first and second genetic elements, and the third identity-function (third row-vector) maps to the open set composed of the second and third genetic elements.

Multiple π mappings

The π mapping represented in **Figure 1** and matrix ***H*** is just one of the possible mappings from the set *F* to the topological space (G, τ). In general, given a *F* set with m identify-functions, and a genetic set with n genetic elements, standard set theory shows that the topological space (G, τ) has 2^n open sets resulting in $\frac{2^n!}{(2^n-m)!}$ possible injective π mappings from *F* to (G, τ). Although it is possible that many, if not most, of these potential mappings do not exist due to negative natural selection, there is still a big reservoir of mappings available to the organism. We will use the notation $\boldsymbol{F} \overset{\boldsymbol{\pi_j}}{\to} (\boldsymbol{G}, \boldsymbol{\tau})$**,** where π_j is the j-th mapping from F to (G, τ), to represent different mappings. Each π_j mapping can be represented by an m x n $\boldsymbol{H_j}$ matrix as described above. The row-vector $\boldsymbol{r_i^j}$ denotes the open set $\pi_j(f_i)$ that is, the set of genetic elements mapped by the i-th identity-function in the π_j mapping. We will call Π as the set of possible mappings of an organism: $\Pi = \{\pi_1, \pi_2, \ldots, \pi_p\}$.

Similarity between π mappings

Let us to define and calculate a similarity metric – or its opposite, a distance metric – between mappings. We will define the similarity *S* between mappings π_i and π_j as the average similarity between open sets $\pi_i(f_k)$ and $\pi_j(f_k)$, *k* = 1, 2, …, m where m equals the number of identity-functions. Similarity between open sets $\pi_i(f_k)$ and $\pi_j(f_k)$ is the number of shared genetic elements (i.e. intersection) divided by the total number of genetic elements in both open sets (i.e. union):

$$S(\pi_i, \pi_j) = \frac{1}{m}\sum_{k=1}^{m} S[\pi_i(f_k), \pi_j(f_k)] = \frac{1}{m}\sum_{k=1}^{m} \frac{|\pi_i(f_k) \cap \pi_j(f_k)|}{|\pi_i(f_k) \cup \pi_j(f_k)|}; \qquad 0 \le S(\pi_i, \pi_j) \le 1$$

A distance metric between mappings π_i and π_j can be defined as

$$D(\pi_i, \pi_j) = 1 - S(\pi_i, \pi_j); \qquad 0 \le D(\pi_i, \pi_j) \le 1$$

As example, three different mappings, with three identity-functions and three genetic elements, are shown below

$$\boldsymbol{F} \overset{\boldsymbol{\pi_1}}{\rightarrow} (\boldsymbol{G}, \boldsymbol{\tau}) \equiv \ \boldsymbol{H_1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}; \quad \pi_1(f_1) = G_1, \quad \pi_1(f_2) = G_2, \quad \pi_1(f_3) = G_3$$

$$\boldsymbol{F} \overset{\boldsymbol{\pi_2}}{\rightarrow} (\boldsymbol{G}, \boldsymbol{\tau}) \equiv \ \boldsymbol{H_2} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}; \quad \pi_2(f_1) = G_1, \quad \pi_2(f_2) = G_3, \quad \pi_2(f_3) = G_2$$

$$\boldsymbol{F} \overset{\boldsymbol{\pi_3}}{\rightarrow} (\boldsymbol{G}, \boldsymbol{\tau}) \equiv \; \boldsymbol{H_3} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}; \quad \pi_3(f_1) = G_{12}, \quad \pi_3(f_2) = G_{13}, \quad \pi_3(f_3) = G_{23}$$

Then

$$S(\pi_{1,}\pi_2) = \frac{1}{3}\left(\frac{1}{1} + \frac{0}{2} + \frac{0}{2}\right) = \frac{1}{3}$$

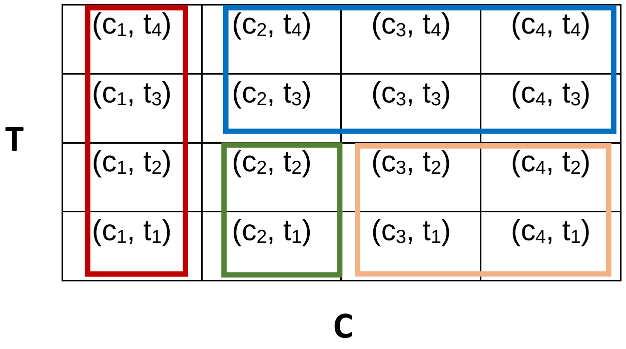
$$S(\pi_{1,}\pi_3) = \frac{1}{3}\left(\frac{1}{2} + \frac{0}{3} + \frac{1}{2}\right) = \frac{1}{3}$$

$$S(\pi_{2,}\pi_3) = \frac{1}{3}\left(\frac{1}{2} + \frac{1}{2} + \frac{1}{2}\right) = \frac{1}{2}$$

The EpG^2 system as a fiber bundle

Individuals consist of different cell types that develop over time. Let us define *C* as the set of all cell types of an organism, *T* as the set of all developmental times, and *S* as the Cartesian product of *C* and *T*, $S = C \times T$. We will show that the set Π (i.e. the set of possible π mappings) generates a partition of *S*. For simplicity –without lack of generality– let us assume four different cell types, *C* = {c_1, $c_{2,}$ c_3, c_4}; four different developmental times, *T* = {t_1, t_2, t_3, t_4}; and four different π mappings, Π = {π_1, π_2, π_3, π_4}. **Figure 2** shows an example of the space *S* and its partition generated by Π.

FIGURE 2



(c_1, t_4)		(c_2, t_4)	(c_3, t_4)	(c_4, t_4)
(c_1, t_3)		(c_2, t_3)	(c_3, t_3)	(c_4, t_3)
(c_1, t_2)		(c_2, t_2)	(c_3, t_2)	(c_4, t_2)
(c_1, t_1)		(c_2, t_1)	(c_3, t_1)	(c_4, t_1)

Figure 2. Space S = C x T of four cell types and four developmental times. The coordinate pair (c_i, t_j) denotes the i-th cell type in the j-th developmental time. Each color represents a different π mapping: π_1 (red), π_2 (blue), π_3 (green), and π_4 (orange). Coordinate pairs under the same color form an equivalence class; they use the same π mapping (i.e. they have the same genome).

The coordinate pair (c_i, t_j) represents the i-th cell type in the j-th developmental time. Two coordinate pairs are equivalent, $(c_i, t_j) \sim (c_k, t_l)$, if they use the same π mapping. In the example shown in **Figure 2**, the equivalence relation generates four equivalence class. The set of equivalence classes generated by the equivalence relation ~ is called the quotient set of **S** by ~ and is denoted by *S*/~. In the present example,

$S/\sim = \{\varphi_1, \varphi_2, \varphi_3, \varphi_4\}$, where

$\varphi_1 = \{(c_1, t_1), (c_1, t_2), (c_1, t_3), (c_1, t_4)\}$

$\varphi_2 = \{(c_2, t_3), (c_2, t_4), (c_3, t_3), (c_3, t_4), (c_4, t_3), (c_4, t_4)\}$

$\varphi_3 = \{(c_2, t_1), (c_2, t_2)\}$

$\varphi_4 = \{(c_3, t_1), (c_3, t_2), (c_4, t_1), (c_4, t_2)\}$

The quotient map θ is the surjective function that sends each member of *S* to its respective equivalence class. Symbolically

$$\boldsymbol{S} \overset{\boldsymbol{\theta}}{\rightarrow} \boldsymbol{S}/\sim$$

The set Π of π mappings and the quotient set *S*/~ of equivalence classes are isomorphic to each other, as there is a one-to-one correspondence between π mappings and φ equivalence classes (see **Figure 2**). Symbolically

$$\boldsymbol{\Pi} \cong \boldsymbol{S}/\sim$$

Because of the isomorphism between Π and *S*/~ we can also say that the quotient function θ sends each element of S to its corresponding π mapping,

$$\boldsymbol{S} \xrightarrow{\boldsymbol{\theta}} \boldsymbol{\Pi}$$

We are ready now to define the EpG^2 system from a set theory perspective. The EpG^2 system is a structure constituted of the space *S* and the space Π. As **Figure 2** shows, the different π mappings generate a partition of the space *S*. Using standard topology terminology, we say that the EpG^2 system is a fiber bundle with base space *S* and fiber Π. In particular, the EpG^2 system may be thought as composed of a group of fibers (i.e. π mappings) "above" the base space *S*. Each particular fiber maps to an open set in the base space *S*.

Genetic basis of complex diseases

Ongoing research tries to find the genetic basis (i.e. DNA sequence variation) of complex human diseases. Some points should be considered based on the hypothesis proposed in the current work:

1) To map a complex phenotypic trait to DNA elements (i.e. the genetic space (G, τ)), we first need a map from the set of phenotypic traits (*P*) to the set of biological identity-functions (*F*).
2) A complex phenotypic trait is a global time-dependent property of the individual. This means, an emergent property of the whole individual.
3) Because of points 1) and 2), the mapping from phenotypic traits to biological identity-functions, and therefore to DNA elements, will vary by developmental time and cell type.

Taking into consideration the above points, let us make the following definitions:

Definition 5

- *P* is the set of *q* phenotypic traits of the organism
- *P* = {p_1, p_2, …, p_q}, where p_i is the i-th phenotypic trait

Definition 6

- (F, τ) is the biological identity-function space with discrete topology, which means every possible subset of F –including the empty set ϕ and F itself are open sets– is open. For example, let us assume a biological identity-function with three elements: F = {f_1, f_2, f_3}. Then, the topological space (F, τ) will include all possible subsets of F: (F, τ) = {ϕ, {f_1}, {f_2}, {f_2}, {f_1, f_2}, {f_1, f_3}, {f_2, f_3}, {f_1, f_2, f_3}}.
- We will use the notation $F_{ij\ldots k}$ to represent the {f_i, f_j, …, f_k} open set. According to this notation, the above topological space (F, τ) can be written as (F, τ) = {ϕ, F_1, F_2, F_3, F_{12}, F_{13}, F_{23}, F_{123}}.

Definition 7

- $\sigma: P \rightarrow (F, \tau)$ is an injective mapping from the *P* set to the topological space (*F*, τ) that is, if two phenotypic traits p_a and p_b map to the same open set $F_{ij\ldots k}$ then p_a and p_b are the same phenotypic trait. In other words, for a given σ mapping two different phenotypic traits cannot map to the same open set. Symbolically, $\forall p_a, p_b \in P$, $\sigma(p_a) = \sigma(p_b) \Rightarrow p_a = p_b$.
- Assuming a *P* set with two phenotypic traits and a F set with three identity function, **Figure 3** shows an example of σ mapping from the *P* set to the (F, τ) identity-function space.
- The σ mapping shown in **Figure 3** is just one of multiple mappings that will vary by cell type and developmental time. In general, $\sigma_{ij}: P \rightarrow (F, \tau)$ will be the σ mapping in the i-th cell type and j-th developmental time.

FIGURE 3

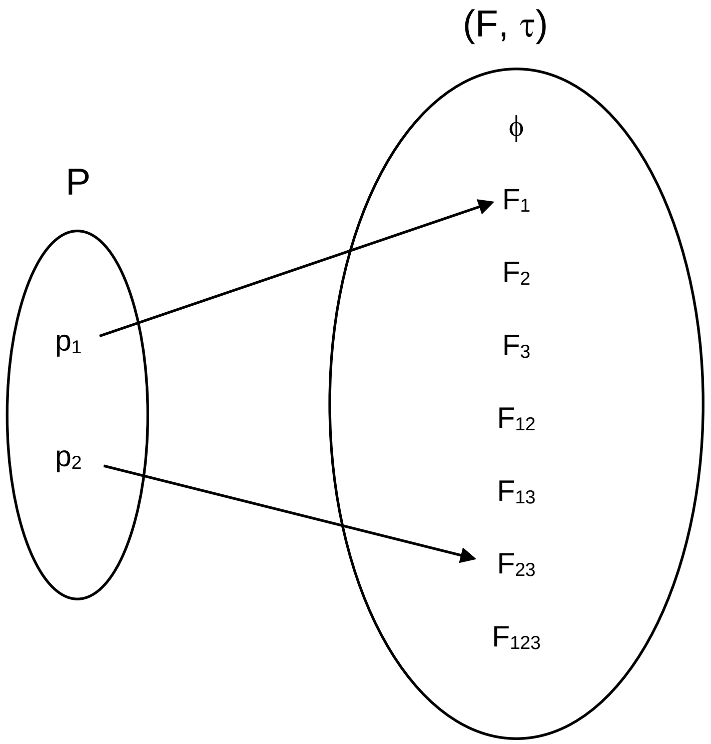


Figure 3. An example of σ mapping of the set *P* of phenotypic traits to the topological space of identity-functions (*F*, τ). In particular, σ(p_1) = F_1, and σ(p_2) = F_{23}.

Definition 8

- The $\gamma: P \rightarrow (G, \tau)$ mapping from the *P* set to the genetic topological space (*G*, τ) is the combination of the $\pi: F \rightarrow (G, \tau)$ and $\sigma: P \rightarrow (F, \tau)$ mappings as follows

$$\begin{tikzcd}
F \arrow[r, "\pi"] & G \\
P \arrow[u, "\sigma"] \arrow[ur, "\gamma"'] &
\end{tikzcd}$$

$\gamma = \pi \circ \sigma$ = π after σ

- For example, combining the π mapping from **Figure 1**: $\pi(f_1) = G_2$, $\pi(f_2) = G_{12}$, and $\pi(f_3) = G_{23}$; with σ mapping from **Figure 3**: $\sigma(p_1) = F_1$, and $\sigma(p_2) = F_{23}$, we get the corresponding γ mapping: $\gamma(p_1) = G_2$, and $\gamma(p_2) = G_{123}$.
- The γ mapping will vary over cell type and developmental time. In general, $\gamma_{ij}: P \to (G, \tau) = \pi_{ij} \circ \sigma_{ij}$ will be the γ mapping in the i-th cell type and j-th developmental time.

Definition 9

- A phenotypic trait including diseases are global properties of the organism. This means, a particular phenotypic trait *P_A* at time *T*, which we will call P_A^T, depends on the life history of the individual organism up to time *T* and is an emergent property of the whole organism.
- $P_A^T = I_i^C I_j^T p_{aij}$ will be the particular phenotypic trait *P_A* at time *T*. The $\boldsymbol{I}$ operator represents a biological integration over cell type and developmental time. In particular:

  I_i^C = biological integration over all *C* cell types (i = 1, 2, …, C),

  I_j^T= biological integration up to time *T* (j = 1, 2, …, T),

  p_{aij}= contribution of the i-th cell type at the j-th developmental time to the phenotypic trait *P_A*. In other words, p_{aij} is the set of identity-functions in the i-th cell type at the j-th developmental time that contributes to the global phenotypic trait *P_A*.

DISCUSSION AND IMPLICATIONS

The genome and epigenome do not exist independent from each other

A major claim of the present work is that the genome and the epigenome do not have independent existence from each other. Present-day paradigm assigns ontological primacy to the genome over the epigenome, that is, the former exists –in an ontological sense– before the latter. For example, a survey of textbooks, scientific articles, websites, etc. shows that the genome is defined without any reference to the epigenome, which at the same time is defined always in reference to the genome. Current thinking is based on the identification of an organism's genome with the whole of its DNA, and the epigenome with the complete set of chemical changes on the DNA itself (e.g. cytosine methylation) or associated proteins (e.g. histone modifications) regulating genome's activity. However, as I argue in the present work, the genome and the epigenome are different aspects of the same EpG^2 system. The π mapping illuminates these two aspects. For example, we could define an organism's genome as the image (i.e. range) of the π mapping that is, the set of DNA elements that are the output of the π mapping. However, we should notice that a DNA element g_A becomes a genomic element only after being paired with their respective identity-function f_A. In other words, it is the ordered pair (f_A, g_A) that is a genomic element rather than the DNA element g_A alone. Therefore, we can define the genome as the set of all ordered pairs (f, g) for a particular π mapping; which for definition, based on standard set theory, is the π mapping itself. We should also note that because the same individual organism has several different π mappings (i.e. fibers of the EpG^2 system) then, it would have several different genomes too. This multiplicity of genomes in the same individual should not be confused with somatic differences of DNA sequence across tissues (Yizhak et al., 2019). Rather, the present work proposes that DNA sequence is not synonymous of genome, and different genomes can have the same underlying DNA sequence.

On the other hand, how can we define the epigenome? Based on current definitions, the epigenome consists of chemical changes to the DNA molecule (e.g. cytosine methylation) and associated proteins (e.g. histone modifications) that regulate genome's activity (e.g. gene expression). I postulate that the epigenome, more than just telling the genome what do, defines what the genome is. Then, we can define the epigenome as the set of mechanisms leading to the emergence of a particular π mapping (i.e. genome) in a particular cell type and developmental time.

It is clear from the above discussion that the genome has no independent existence from the mechanisms (i.e. epigenome) resulting on its emergence. In addition, because a mechanism can be understood as a process occurring in a particular system (Bunge, 1997) then then the epigenome exists a part of a system: the EpG^2 system.

A multilevel epigenome

Our definition of the epigenome as the mechanisms responsible for the emergence of the genome can be expanded to include regulatory mechanisms of genome's activity. The epigenome will be then a series of multi-level mechanisms responsible for the emergence of the genome and the regulation of the genome's activity (**Figure 4**). The first, high-level epigenome will control the assignment of function-identity to the different DNA elements. This mapping of function-identity all throughout the DNA molecule will result on the emergence of the genome. Once again, the present work advances the hypothesis that this mapping can vary by development time and cell type (i.e. same DNA elements may have different function-identity across the *S* space, see **Figure 2**). Therefore, the genome is not an absolute entity but rather an emergent one. The second, low-level epigenome will regulate the activity of the emerged genome. For example, levels of gene expression, use of alternative promoters, etc. Unfortunately, because present day paradigm assigns ontological primacy to the genome over the epigenome most, if not all, of ongoing research focuses on the lower level of the epigenome, and there is a complete lack of studies about high-level epigenomics mechanisms leading to the emergence of the genome.

FIGURE 4

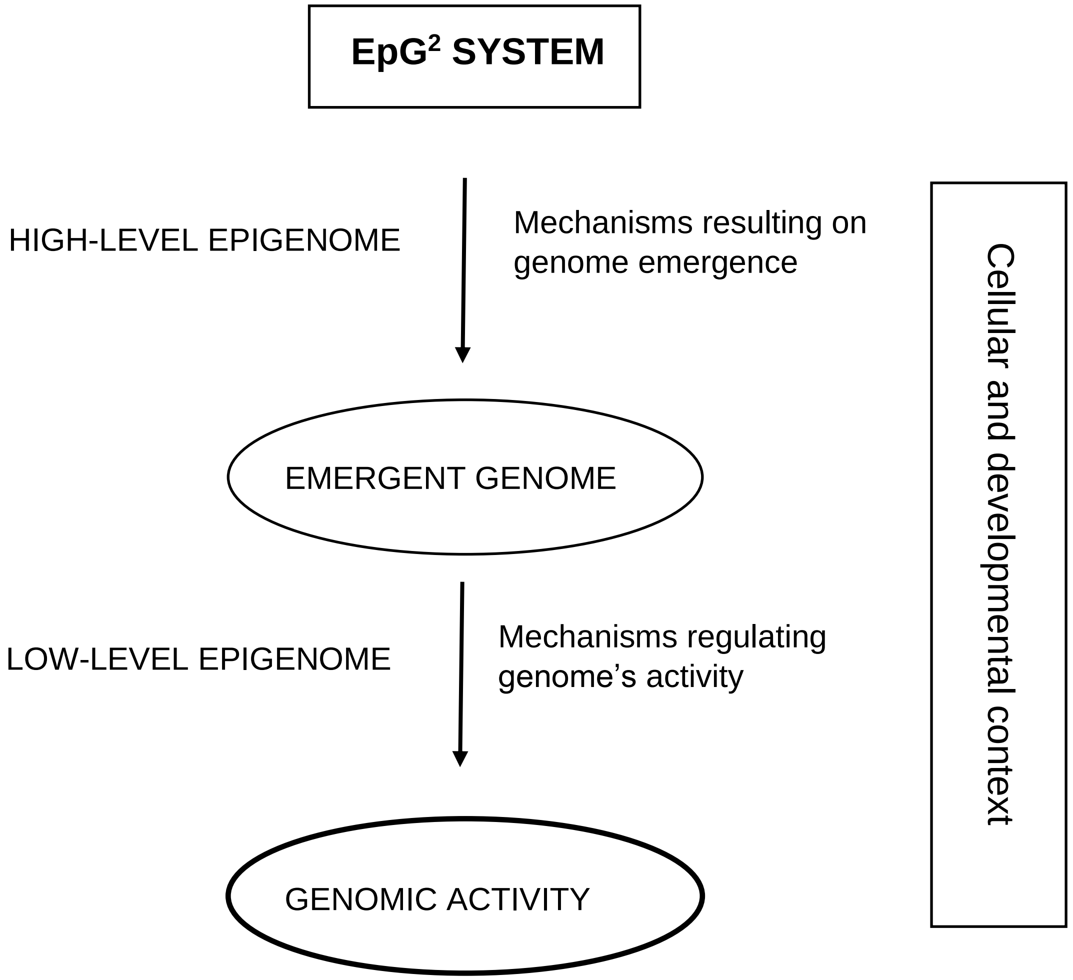

EpG2 SYSTEM
HIGH-LEVEL EPIGENOME
Mechanisms resulting on genome emergence
EMERGENT GENOME
LOW-LEVEL EPIGENOME
Mechanisms regulating genome’s activity
GENOMIC ACTIVITY
Cellular and developmental context

Figure 4. The multi-level epigenome and the emergent genome. The first, high-level epigenome consists of a series of mechanisms resulting on the emergence of the genome. The second, low-level epigenome is responsible for the regulation of genome’s activity. The multi-level epigenome as well as the emergent genome will depend of cellular and developmental time context.

A robust, research program must take into account the multi-level epigenome and the emergent genome. In the following, I suggest some relevant areas of investigation that arise from the proposed theory:

1. **Epigenomics mechanisms leading to the emergence of the genome.** Because current paradigm considers the genome as an absolute entity, whose existence does not depend on epigenomics processes, this question is not even asked. A robust research program should identify high-level epigenomics mechanisms responsible for the emergence of the genome throughout different cell types and developmental times.
2. **Individual variation of the multi-level epigenome.** To date, it is widely recognized the existence of person-to-person variation on low-level epigenomics mechanisms (e.g. DNA methylation) that may result on differences of genome's activity (e.g. levels of gene expression). Our proposed theory posits the existence of a high-level epigenome, which is responsible for genome emergence (**Figure 4**). Future research should address several questions:
    - **a.** Is there inter-individual variation in the first, high-level epigenome as well (i.e. person-to-person variation in the π mappings)?
    - **b.** Factors affecting variation of the multi-level epigenome.
    - **c.** Relationship between variation of the multi-level epigenome and phenotypic traits including disease.
3. **Genomic variation and disease.** An active field of research is the elucidation of the hereditary basis of disease. What this means in practice is the assessment of how DNA sequence variation affects both risk and severity of disease. However, as proposed in the current work, an organism can be viewed as a bundle of different genomes (i.e. π mappings) that vary by cell types and developmental times. In must be noted again, that these different genomes may have the same underlying DNA sequence, as they refer to different π mappings. As consequence, same physical positions along the DNA molecule may have different identity-function depending on the respective genome (i.e. π mapping). In addition, if there is person-to-person variation of the π mappings (see point 2 above) inter-individual

variation in the DNA sequence is not equivalent to inter-individual genomic variation. Standard approaches such as genome-wide association studies and whole-genome sequencing association studies would fail to elucidate how genome variation affects disease, as they focus on variation of the underlying DNA sequence that is not necessarily the same as genomic variation. A new approach should consider genomic variation across cell types and developmental times as the effect of DNA sequence variation on disease cannot be disentangled from the subject's life-story as genomes (see **genetic basis of complex diseases** section).

SUMMARY

In the present work, I argue that even though the DNA is part of the material basis of the genome, the latter is more than the DNA sequence. The genome is rather an emergent entity resulting from epigenomics mechanisms. The proposed hypothesis has important implications for the study of the epigenomics and genomic basis of phenotypic traits, including diseases, and offers a new paradigm for future research.

REFERENCES

Andersson, R., 2015. Promoter or enhancer, what's the difference? Deconstruction of established distinctions and presentation of a unifying model. Bioessays 37, 314-23, doi:10.1002/bies.201400162.

Bunge, M., 1997. Mechanism and Explanation. Philosophy of the Social Sciences 27, 410-465, doi:10.1177/004839319702700402.

Burris, H. H., Baccarelli, A. A., 2014. Environmental epigenetics: from novelty to scientific discipline. J Appl Toxicol 34, 113-6, doi:10.1002/jat.2904.

Carlson, E. A., 1991. Defining the gene: an evolving concept. Am J Hum Genet 49, 475-87.

Domene, S., Bumaschny, V. F., de Souza, F. S., Franchini, L. F., Nasif, S., Low, M. J., Rubinstein, M., 2013. Enhancer turnover and conserved regulatory function in vertebrate evolution. Philos Trans R Soc Lond B Biol Sci 368, 20130027, doi:10.1098/rstb.2013.0027.

Erceg, J., Pakozdi, T., Marco-Ferreres, R., Ghavi-Helm, Y., Girardot, C., Bracken, A. P., Furlong, E. E., 2017. Dual functionality of cis-regulatory elements as developmental enhancers and Polycomb response elements. Genes Dev 31, 590-602, doi:10.1101/gad.292870.116.

Fourel, G., Magdinier, F., Gilson, E., 2004. Insulator dynamics and the setting of chromatin domains. Bioessays 26, 523-32, doi:10.1002/bies.20028.

Gerstein, M. B., Bruce, C., Rozowsky, J. S., Zheng, D., Du, J., Korbel, J. O., Emanuelsson, O., Zhang, Z. D., Weissman, S., Snyder, M., 2007. What is a gene, post-ENCODE? History and updated definition. Genome Res 17, 669-81, doi:10.1101/gr.6339607.

Griffiths, P. E., Stotz, K., 2006. Genes in the postgenomic era. Theor Med Bioeth 27, 499-521, doi:10.1007/s11017-006-9020-y.

Hnisz, D., Abraham, B. J., Lee, T. I., Lau, A., Saint-Andre, V., Sigova, A. A., Hoke, H. A., Young, R. A., 2013. Super-enhancers in the control of cell identity and disease. Cell 155, 934-47, doi:10.1016/j.cell.2013.09.053.

Kim, T. K., Shiekhattar, R., 2015. Architectural and Functional Commonalities between Enhancers and Promoters. Cell 162, 948-59, doi:10.1016/j.cell.2015.08.008.

Kolovos, P., Knoch, T. A., Grosveld, F. G., Cook, P. R., Papantonis, A., 2012. Enhancers and silencers: an integrated and simple model for their function. Epigenetics Chromatin 5, 1, doi:10.1186/1756-8935-5-1.

Maston, G. A., Evans, S. K., Green, M. R., 2006. Transcriptional regulatory elements in the human genome. Annu Rev Genomics Hum Genet 7, 29-59, doi:10.1146/annurev.genom.7.080505.115623.

Palstra, R. J., Grosveld, F., 2012. Transcription factor binding at enhancers: shaping a genomic regulatory landscape in flux. Front Genet 3, 195, doi:10.3389/fgene.2012.00195.

Pesole, G., 2008. What is a gene? An updated operational definition. Gene 417, 1-4, doi:10.1016/j.gene.2008.03.010.

Portin, P., Wilkins, A., 2017. The Evolving Definition of the Term "Gene". Genetics 205, 1353-1364, doi:10.1534/genetics.116.196956.

Raab, J. R., Chiu, J., Zhu, J., Katzman, S., Kurukuti, S., Wade, P. A., Haussler, D., Kamakaka, R. T., 2012. Human tRNA genes function as chromatin insulators. EMBO J 31, 330-50, doi:10.1038/emboj.2011.406.

Rebeiz, M., Tsiantis, M., 2017. Enhancer evolution and the origins of morphological novelty. Curr Opin Genet Dev 45, 115-123, doi:10.1016/j.gde.2017.04.006.

Scherrer, K., Jost, J., 2007. Gene and genon concept: coding versus regulation. A conceptual and information-theoretic analysis of genetic storage and expression in the light of modern molecular biology. Theory Biosci 126, 65-113, doi:10.1007/s12064-007-0012-x.

Shin, H. Y., 2018. Targeting Super-Enhancers for Disease Treatment and Diagnosis. Mol Cells 41, 506-514, doi:10.14348/molcells.2018.2297.

Van Bortle, K., Corces, V. G., 2012. tDNA insulators and the emerging role of TFIIIC in genome organization. Transcription 3, 277-84, doi:10.4161/trns.21579.

Villar, D., Berthelot, C., Aldridge, S., Rayner, T. F., Lukk, M., Pignatelli, M., Park, T. J., Deaville, R., Erichsen, J. T., Jasinska, A. J., Turner, J. M., Bertelsen, M. F., Murchison, E. P., Flicek, P., Odom, D. T., 2015. Enhancer evolution across 20 mammalian species. Cell 160, 554-66, doi:10.1016/j.cell.2015.01.006.

Wei, W., Brennan, M. D., 2001. The gypsy insulator can act as a promoter-specific transcriptional stimulator. Mol Cell Biol 21, 7714-20, doi:10.1128/MCB.21.22.7714-7720.2001.

Yang, S., Oksenberg, N., Takayama, S., Heo, S. J., Poliakov, A., Ahituv, N., Dubchak, I., Boffelli, D., 2015. Functionally conserved enhancers with divergent sequences in distant vertebrates. BMC Genomics 16, 882, doi:10.1186/s12864-015-2070-7.

Yizhak, K., Aguet, F., Kim, J., Hess, J. M., Kübler, K., Grimsby, J., Frazer, R., Zhang, H., Haradhvala, N. J., Rosebrock, D., Livitz, D., Li, X., Arich-Landkof, E., Shoresh, N., Stewart, C., Segrè, A. V., Branton, P. A., Polak, P., Ardlie, K. G., Getz, G., 2019. RNA sequence analysis reveals macroscopic somatic clonal expansion across normal tissues. Science 364, doi:10.1126/science.aaw0726.