Structure Encoding in DNA

Antony Van der Mude

Burlington MA 01803

Corresponding author:
Antony Van der Mude

Email address: vandermude@acm.org

ABSTRACT

It is proposed that transposons and related long non–coding RNA define the fine structure of body parts. Although morphogens have long been known to direct the formation of many gross structures in early embryonic development, they do not have the necessary precision to define a structure down to the individual cellular level. Using the distinction between procedural and declarative knowledge in information processing as an analogy, it is hypothesized that DNA encodes fine structure in a manner that is different from the genetic code for proteins. The hypothesis states that repeated or near–repeated sequences that are in transposons and non–coding RNA define body part structures. As the cells in a body part go through the epigenetic process of differentiation, the action of methylation serves to inactivate all but the relevant structure definitions and some associated cell type genes. The transposons left active will then physically modify the DNA sequence in the heterochromatin to establish the local context in the three–dimensional body part structure. This brings the encoded definition of the cell type to the histone. The histone code for that cell type starts the regulatory cascade that turns on the genes associated with that particular type of cell, transforming it from a multipotent cell to a fully differentiated cell. This mechanism creates structures in the musculoskeletal system, the organs of the body, the major parts of the brain, and other systems.

INTRODUCTION

As currently understood, the primary purpose of DeoxyriboNucleic Acid (*DNA*) in the cell is for long–term information storage of the specification for proteins, where the synthesis of proteins is controlled by RiboNucleic Acid (*RNA*). It is usually considered that: "DNA makes RNA and RNA makes protein" (Wikipedia contributors, 2018a). This is the process of gene expression. A gene is "a sequence of DNA or RNA that codes for a molecule that has a function." (Wikipedia contributors, 2018b). This is the heart of the Central Dogma of Molecular Biology, the formulation of the general rules for information transfer in genetics (Crick, 1970).

But a large fraction of the DNA does not code for genes. Some non–coding sequences, such as introns, are important for gene regulation. But there are large sequences in the intergenic regions that had previously been assumed not to have a purpose. Many of these regions contain a high degree of repetitive elements. It is likely that there are other independent processes of information transfer at work that are different from the coding of proteins. One possibility is an encoding for structure.

Certain genes determine the structure of a multicellular organism such as the homeodomain proteins. Concentration of various morphogens lay out the somatic structure (Lawrence and Struhl, 1996). But they seem to only define the gross structure of the parts that make up the organism. Determination of the fine structure requires something more.

Kerszberg and Wolpert (Kerszberg and Wolpert, 2007) point out some of the problems. They note that morphogens may lack the necessary precision and robustness to determine positional information down to the cellular level. But it appears that this precision is needed to specify the fine structure of a body part. There is a lot of noise in the molecular concentrations of morphogens, for example, making it difficult for morphogens alone to determine the fine structure.

> If morphogens are to act as graded positional cues, then there must exist mechanisms for cells to perceive and interpret concentration–dependent information, and this raises problems. For example, if position is specified on a cell–by–cell basis then many more morphogen

> concentration thresholds (at which changes in gene activity occur) need to be established than the five or so identified in some tissues. ... We suggest that, just like the mechanisms involved in polarization and somite formation, those for setting up positional values may involve cell–cell interactions. There might even exist an overlap among the molecular players in these seemingly independent sets of phenomena. Morphogenetic molecules do exist, but it seems improbable that their concentration alone determines the fate of cells regarding their final position in the developing embryo. Wardle and Smith (Wardle and Smith, 2004) reported that early in development gene expression at the single–cell level is rather variable and only later does it become more precisely linked to cell position. Thus, morphogens may represent a rather crude positional information system, which is then more finely tuned by cell–cell interactions. Clearly, the morphogen gradient does not act alone and is itself specified by a variety of complex cellular mechanisms.

The human genome contains tens of thousands of genes, far less than what is needed to define the detailed structure of all the body's parts. Some studies have shown that most of the genome does not code for proteins. Of the genes that actually do code for proteins, current studies show that there are only about 20,000 of them, and many are among the oldest and most conserved sequences (Ezkurdia et al., 2014).

We hypothesize that there is additional structure information in the DNA that is kept in the intergenic regions, in its own unique encoding. This encoding is most likely unrelated to the codons that determine protein structure. This hypothesis is based on an analogy to computer memory storage using information–theoretic arguments. It is suggested that this structure information would appear as non–coding sequences.

Here are some examples of the fine structure that would be defined by a structure encoding (Gray, 1977):

- The human femur is composed of a number of parts, including head, trochanter, and condyles. These parts all have a specific structure.
- The heart is separated into a number of gross structures, such as the aorta and the atria and ventricles, that follow a specific developmental sequence. But there are a number of subparts that have a detailed structure. For example, the valves (mitral, tricuspid, pulmonary and aortic) are all shaped differently. They appear later in development.
- The brain is composed of discrete parts, such as the amygdala or the hippocampus. For these parts to perform their function, there must be specific structure information, possibly even down to the individual neuron level.

In this paper, we will use the term *body part* to refer to a collection of cells that form a structure. A body part can be a collection of cells composing a part of the musculature (myocytes and adipocytes), skeleton (osteoblast or chondrocytes), brain (neurons and glial cells) or the cells of the other organs of the body. Since the discussion involves fine structure, some of these parts could be small substructures of a larger system or somite, such as the glomerulus in the kidney or the alveolus in the lungs. Since the focus is on the fine structure of body parts large and small, the usage of the term body part is made without distinction to the size of the part or the larger organ it resides in. We will reserve the term *structure* (structure information) to refer to the detailed encoding of a body part.

ANALOGY: THE COMPUTER HARD DRIVE

Computers are composed of a processor, input/output devices and memory. The memory is either volatile or static. Static memory fills a similar purpose for a computer that DNA does for a cell. The static memory of a computer (the hard drive or flash memory) contains a variety of encoded data in permanent storage. There are the programs that control the operating system. There are applications that are run on the computer. There is also data that is used by the applications, such as documents and pictures.

The data in static memory can be said to be composed of two types: procedural and declarative knowledge (Winograd, 1975).

Procedural knowledge includes the commands of the operating system (copy file, rename directory, print text, draw window on the screen). The commands are analogous to the proteins in a cell. They do something. This is also true for the application programs. But data such as documents and pictures do

not control the operation of the computer, yet they usually comprise the largest part of the information encoded on a computer hard drive. This other data is the declarative knowledge.

The declarative data is read by the application programs, which then use the commands of the operating system to do things, based on that data. For instance, a picture can be copied, edited, deleted, transmitted or printed.

It is important to note that if a user desires to store a document or picture, the data that represents that item is not stored as part of the application program or part of the operating system. It is stored by itself in its own encoding.

A simple assembly language example will show the different types of computing data (Noordergraaf and Boldyshev, 2013):

```
section     .text
global      _start                  ;must be declared for linker (ld)
_start:                             ;tell linker entry point
    mov     edx,len                 ;message length
    mov     ecx,msg                 ;message to write
    mov     ebx,1                   ;file descriptor (stdout)
    mov     eax,4                   ;system call number (sys_write)
    int     0x80                    ;call kernel
    mov     eax,1                   ;system call number (sys_exit)
    int     0x80                    ;call kernel
section     .data
msg     db  'Hello, world!',0xa     ;data string to output
len     equ $ - msg                 ;length of data string
```

In this example, there is a distinction between program (many assemblers refer to the program code as "text") and data — the procedural and declarative knowledge. Although in this case there are two separate sections, there is nothing to prohibit interspersing code and data if desired. This is certainly true about the way programs and data are stored in static memory. Like DNA, the individual elements are not usually required to be all in a particular order — except for reasons of efficiency.

The encoding of the computer program is in machine language. This is analogous to the genetic code that defines the sequence of amino acids that make up a protein. In this example, there are only two types of instructions, one to move data from one place to another and the other to call system subroutines. Other machine language instructions can do things like perform arithmetic operations or make comparisons of two pieces of data.

The encoding of the declarative data is completely different from the encoding of the program. In this example, the dataset representing the output string is encoded in ASCII (American Standard Code for Information Interchange), where the letter "H" is encoded as 72, "e" is 101, "l" is 108, "o" is 111, and so on. The code "0xa" is a Carriage Return — the end of the output line.

Pictures are often encoded as an array of pixels, each of which is a point that has a particular color. A dataset for a picture almost always begins with some header information, such as the width and height of the image in pixels. Then it is possible to determine which row and which column a particular pixel represents. Note that this is not the only way to store two or even three dimensional data.

Analogous to the dichotomy of procedural and declarative knowledge in a computer's static memory, we shall argue that there are at least two encodings in cellular DNA: the instructions to make the proteins that are the machines of the cell, and also an encoding of structure information that lays out in detail the positions of the types of cells in the body.

This is the basis of the hypothesis. There are some other important observations from computer science that will further suggest how the structure information is stored and used.

First, many applications make a distinction between the overall gross structure of an item and the details. For example, a document processor may have templates for letters, books, and essays stored as part of the application program, but the details of each individual document are kept separately. Similarly, the cell may have proteins and chemical signals that define the gross structure of a somite, but the fine structure could be kept as a separate encoding. It may also be true that different encodings could be used for different templates. In any case, the process goes as follows: a top–level template is chosen, then perhaps a more detailed template is applied to that, and so on. At that point, the declarative information is

used to arrive at the fine structure. This fine structure identifies which character (for a document) or pixel (for an image), leading to a final fixed determination for that datum.

Second, although many computer systems use data compression to efficiently store information in static memory, most of these techniques do not tolerate data processing errors very well. Compression algorithms are often based on patterns that are found in a dataset taken as a whole. Therefore errors introduced at one point in the compressed data can cause problems all through the dataset, even resulting in the dataset being unreadable. In contrast, uncompressed datasets like images have a lot of repetitive elements — that is, sequences that repeat the same element over and over, but with some larger variations. In an image, different color fields would have the same or similar elements, which could then change into different elements in another section of the image. Leaving the data uncompressed is error–tolerant. A few elements could be compromised, but that does not destroy the whole structure.

Third, determining which character comes after the previous character in a text document is straightforward, since the characters are a one–dimensional data sequence. But for two and three–dimensional structures, it is much harder to determine the neighbors of a given pixel. The simplest thing for a computer to do is to determine this arithmetically: for a three dimensional representation that has length L, width W and height H pixels, the pixel at location $< l, w, h >$ is found at position $l + (L * w) + (L * W * h)$. But cells don't work arithmetically. They have to use a different process.

This is as far as the analogy goes. Biological cells are not computers, and proteins function differently from operating system commands. It may be possible to specify a body part down to the cellular level, but the way this information is transmitted from cell to cell, and the implementation of positional information in the cell, is almost certainly different from how it is done in a computer. Cells are biochemical machines, and as such, structure determination is probably implemented by physical manipulation, instead of looking up indexed values in a data array.

HYPOTHESIS: DNA ENCODES BODY PART STRUCTURES

Using the concept of procedural and declarative information in a computer's static memory as an analogy, we shall construct a hypothesis of how detailed structure information is encoded in the DNA and how it is involved in the embryonic development of multicellular organisms.

There are a number of parts to structure encoding:

- Storage of the structure information.
- Determination of gross structure.
- Determination of fine structure.
- Intercellular inheritance and communication of part and location.
- Implementation of structure determination.

This leads to the following questions:

- Where is the structure information stored?
- How to determine which structure the cell is in?
- How is the structure information laid out?
- Does the cell acquire structure information from mitosis?
- How is structure information passed from one cell to another?
- How is the cell type determined from the structure information?
- Once determined, how does the cell differentiate according to its type?
- Are there diseases that are caused by errors in structure?
- How does structure fit into the larger picture of evolution?

We will address each of these questions in turn.

Location of Structure Information

The first part of this hypothesis is that DNA contains encoded structure information. This information does not use the genetic code, which is procedural information. Therefore, it is most likely that the structure sequences are found in the intergenic regions. Also, the length of the encoding does not necessarily have to be triples of nucleotides that are found in the genetic code.

We should expect that structure data will consist of repeated patterns of the same short codes representing a particular type of cell if this organ is mostly a mass of the same types of cells. There may be what appears to be random changes or slowly varying changes. These will depend on the fine structure of the body part. It has been noted that non–coding DNA contains repetitive correlations that protein–coding genes do not (Buldyrev et al., 1995).

One common feature found in DNA is the existence of transposable elements (Bourque et al., 2018). They have the ability to modify the genome by splicing in information. They come in two major classes, retrotransposons and DNA transposons. For the purposes of this paper, we will not make a distinction between them: We will use the term *transposon* to refer to either transposable element. Many transposons contain repeated elements.

The transposons are associated with long non–coding RNA (*lnc–RNA*) (Kapusta et al., 2013) (Kapusta and Feschotte, 2014). The hypothesis states that the lnc–RNA contains structure information which is manipulated by the transposon to derive the exact position information for a given cell (Mattick, 2003) (Kapranov and St Laurent, 2012). It has long been speculated that transposons are not "junk" but are an exaptation (Brosius, 1991) (Brosius and Gould, 1992). This view has been changing – recent studies have shown that lnc–RNA can act as an enhancer that affects transcription (Ørom and Shiekhattar, 2011) (Chen et al., 2017). Instead of being considered parasitic (Palazzo and Lee, 2015), we hypothesize that these elements of the DNA are part of the structural toolkit (Thompson et al., 2016).

It is most likely that each terminally differentiated cell is a part of only one structure. Therefore, there must be a mechanism to turn off the effect of the other structures. This is accomplished through deactivation by methylation. There are tens of thousands of CpG sites in the intergenic regions, and most of them are methylated and thus deactivated (Ikeda and Nishimura, 2015). In many cases, this turns off an associated transposon. It has been noted that, in some organisms, non–CpG methylation is substantially absent from genes, whereas methylation in all contexts is abundant in transposons (Zemach et al., 2010)

Here are some features of transposons that suggest they encode structure information (Bourque et al., 2018):

- Transposons are major components of thousands of lnc–RNAs. These transposons appear more often in the intergenic regions (Kapusta et al., 2013).
- Transposons include repetitive elements and are themselves repeated (Negre and Simpson, 2013).
- Transposons are active in somatic cells in many organisms (Kazazian, 2011).
- Non–coding RNA appears to play important roles in the maintenance of stem cell pluripotency and other developmental processes (Durruthy-Durruthy et al., 2016).
- Transposons are associated with the regulatory networks that control gene expression (Jacques et al., 2013).
- Transposons are involved in epigenetic control via changes to the chromatin (Rebollo et al., 2011).
- Transposon insertions have been associated with human diseases, such as cancer and autoimmune diseases (Hancks and Kazazian, 2016) (Mattick, 2009).

Hox Genes Control Gross Structure

In the computer memory analogy, Hox genes are like templates. They determine the body part that the structure is located in. Assuming transposons control the structure encoding, the hypothesis states that the Hox genes control the gross structure, but also control which fine structure sequence to use by selecting the transposons and lnc–RNAs.

Homeobox proteins are known to be involved in altering the shape of DNA (Bürglin and Affolter, 2016) by chromatin remodeling (Iimura and Pourquié, 2007). This could be part of the process of structure determination.

Some transposons and lnc–RNAs occur in the Hox clusters (Rinn et al., 2007) (Lempradl and Ringrose, 2008) (Tsumagari et al., 2013) (Feiner, 2016). It is possible that the transposons in Hox gene clusters function as a top–level structure encoding of the body part that the Hox gene controls, with other transposons encoding a hierarchy of finer structures. Studies of P–element transposons in the *Drosophila* bithorax complex Hox cluster show that they control the pattern of gene expression (Maeda and Karch, 2009).

At the same time that structure determination is going on, there is a process of gene expression related to cell differentiation (Lim and Maher, 2010). That is, as the body part is being selected, families of genes are turned on that are common to the correct functioning of these families of cell types.

Layout of the Structure Information

There are more cells in the body that there are cell identifiers in the intergenic regions. This would mean that the gross structure of an organ can be specified at one level, but substructures are made up of repeating patterns that are specified just once, as in a computer subroutine. This obviously applies to substructures such as the glomerulus or the alveolus.

The structure information, as well as cell differentiation, is most likely organized in a hierarchy. As noted above, the transposons and lnc–RNA located near the Hox genes define the gross structure. This structure information references more detailed subpart information in the intergenic regions. A large fraction of the human genome consists of short interspersed nuclear elements (*SINEs*) and long interspersed nuclear elements (*LINEs*) (Kapusta et al., 2013). For example, the most common transposable element in humans is the LINE–1 sequence consisting as much as 17% of the human genome (Cordaux and Batzer, 2009). These could be considered as a motif or a superfine structure.

At the top levels of the hierarchy, cell typing information is organized in what are termed Topologically Associating Domains (*TADs*). TADs are local chromatin interaction domains that are highly self–interacting regions. The internal folding and interaction patterns of TADs are highly cell type–specific (Dekker and Heard, 2015). They are found in the Hox gene locus (Dixon et al., 2012) and are controlled by enhancers in the flanking non–coding regions (De Laat and Duboule, 2013). It has also been found that the effects of the enhancers are altered by the three–dimensional chromatin structure of the TAD (Kragesteen et al., 2018). It has also been found that SINEs near the TAD boundaries are involved in controlling TADs (Pope et al., 2014).

Epigenetic Definition of Gross Structure

In embryonic development, the process of epigenetics is used to pass on structure information about the body part that a new cell is to be part of (Felsenfeld, 2014) (Almouzni and Cedar, 2016). It may also provide some gross information about cell location in the structure. Transposons are known to be epigenetically controlled (Sundaram et al., 2014).

Cell differentiation starts by determining which body part a cell is located in. Methylation appears to play an important role in this determination (Jin et al., 2011). In the early stages of embryonic development, there is extensive reprogramming of methylation status (Messerschmidt et al., 2014). The hypothesis states that methylation is the way that a cell is assigned to a body part. Once that is determined, the cell type is further specified by its unique location in the body part, and thus which genes are expressed by that type of cell. This is why methylation does not necessarily need to regulate the genes directly (Walter, 2015).

Epigenesis also involves the modification of chromatin structure. Epigenetic modifications include alterations of the histone tails and chromatin remodeling such as the establishment and preservation of heterochromatin regions (Felsenfeld and Groudine, 2003) (Lim and Maher, 2010) (Jin et al., 2011) (Sundaram et al., 2014). Preservation of the current status of the cell after mitosis is called *bookmarking* (Sarge and Park-Sarge, 2005). It is likely that bookmarking also preserves structure information.

The process of gene regulation proceeds in parallel with structure determination. Recent studies have revealed the folding of the cell–type–specific chromatin structure into TADs. Also, TAD boundaries form within regions of the same epigenetic state (Mateo et al., 2019). Later fine structure information modifies the expression of genes in a TAD for the specific cell type.

Intercellular Communications

This is the difficult part of implementing a structure encoding.

During the process of mitosis, the methylation information that determines the gross structure that the cell is a part of is transferred to the daughter cells. Therefore, the daughter cells are part of the same gross structure. It is possible that, within a body part, some determination of location can be made by epigenetics. But there has to be a separate mechanism to specify which cell type a daughter cell is, in terms of the local context.

Cells need to transmit to their neighbors the relative location of the cell in the structure. It is not likely that this information is a morphogenic concentration. Concentration gradients used to define the gross structure are not sufficient to determine the fine structure (Kerszberg and Wolpert, 2007). It is more likely that positional information is passed using a data sequence, such as RNA transfer.

The problem is that positional information is three dimensional. To precisely form a body part, it is necessary to determine the three–dimensional coordinates. But the structure encoding is stored one–dimensionally, as part of the DNA sequence. The transposons function not only to encode the structure information but also to process the relative location of cells.

This hypothesis will not make any final determination about how this information is transferred. The actual method can be determined by a detailed analysis of how the fine structure information is organized and applied during the process of cell typing.

One possibility is extracellular vesicles (*exosomes*) as a mechanism of cell–to–cell communication. Since exosomes contain defined patterns of messenger RNA, microRNA, lnc–RNA, and occasionally genomic DNA (Tetta et al., 2013), they may transfer genetic information which results in cell type determination in recipient cells. It has been observed that different cells transfer different lnc–RNAs through exosomes (Dragomir et al., 2018).

Some evidence for the role of non–coding RNA in cell development is the relationship between micro–RNA derived from intergenic regions (Mattick, 2003) (Piriyapongsa et al., 2007) and generated by transposons (McCue and Slotkin, 2012). This could be the way that local context is passed between cells.

Selecting the Entry in the Structure

Selection of the specific cell type at that specific location is done through a process of histone modification and transposon splicing. Splicing physically transforms the DNA structure information to determine the local fine structure. Since the cell cannot do arithmetic computations in $< x, y, z >$ coordinates, transposons were exapted to make the determination of position by physically manipulating the DNA.

Morphogenic processes work in tandem with transposons for chromatin remodeling. Hox genes and their associated transposons establish the top–level structure through methylation (Lim and Maher, 2010). Transposons alter the DNA sequence which changes the location of the histone (Rinn et al., 2007). This sets up the structure encoding in the local context. Transferring information about the current cell relative to the neighboring cells results in further manipulation of the chromatin. The transposon adjusts the DNA sequence to bring the correct cell location information to the histone (Kapusta et al., 2013). This, in turn, will determine the cell type. As far as the cell is concerned, if it is a fully differentiated cell, it does not matter to the cell what its three–dimensional coordinates are. All that matters is what cell type it is (or apoptosis for structure boundaries).

The process of cell location is performed by modifications to the DNA through the insertion of transposons. It has been observed that non–coding RNA is associated with alterations in chromatin structure (Rodriguez-Campos and Azorín, 2007) (Khalil et al., 2009) (Yang et al., 2015) (Kobayashi et al., 2017). It has also been noted that this differs with cell type (Cournac et al., 2015). This is why the nucleosome positions in the genome are so specific (Teif et al., 2012). It may be that the Hox gene turns on the particular "beads on a string" that is the structure of interest. The associated non–coding RNA and the data passed through intercellular communications determine which part of the sequence is read by the histone (Tropberger and Schneider, 2013). The histone reads, in the DNA sequence, the individual cell type information for that location. Information in the histone tail adds an extra layer of contextual interpretation.

Note that a non–coding RNA does not have to be fully transcribed to be effective. Instead, position determination in a structure may result from a partial transcription up to the position of the cell in the structure. Once the position of the cell has been determined, then the cell type is identified, and the transcription does not need to continue past that point. This behavior has been noted in LINE–1 insertions (Sun et al., 2018) (Faulkner and Billon, 2018).

Heterochromatin is of interest. It is densely packed and contains highly repetitive sequences (Bannister

and Kouzarides, 2011) (Walter, 2015). Although hetereochromatin has been shown to turn off gene transcription, that may not be its purpose. Instead, it is hypothesized that the purpose of heterochromatin is to control the determination of structure. In support of this, heterochromatin is known to be controlled by transposons (Feschotte C, 2007) (Rebollo et al., 2011). There is a subset of LINE–1 instances that reside in the heterochromatin (Babenko et al., 2017). Also, it has been shown that, although most of heterochromatin transcription is silenced, this silencing is incomplete and associated with histone modifications (Katan-Khaykovich and Struhl, 2005). This could indicate that all but the correct cell type identifier is silenced.

The area where structural determination is made could be in the facultative heterochromatin (Bannister and Kouzarides, 2011). The formation of facultative heterochromatin is known to differ for different cell types.

Three–dimensional studies of the chromosome using chromosome conformation capture methods (e.g., *Hi–C*) have shown that chromosomes are divided into large compartments that contain either active and open (A–compartments) or inactive and closed chromatin (B—compartments) (Lieberman-Aiden et al., 2009) (Dekker and Heard, 2015). A–compartments cluster with other A–compartments, as do B–compartments with B–compartments, correlated with the cell type–specific gene expression (Fortin and Hansen, 2015). Compartments can encompass several directly adjacent TADs that share chromatin state. It is likely that A– compartments are used to regulate genes, and B–compartments are used to determine structure, where these compartments are either euchromation or heterochromatin.

Lamina–associating domains (*LADs*) are gene–poor regions that contain LINEs. They are associated with gene repression when in contact with the nuclear lamina itself (Shevelyov and Nurminsky, 2012) (Van Steensel and Belmont, 2017). Some TADs correspond to LADs (Dekker and Heard, 2015). The contact between LADs and nuclear lamina has been observed to differ between different cell types (Peric-Hupkes et al., 2010). B–compartments are often LADs (Babenko et al., 2017). If heterochromatin is involved in the process of structure determination, LADs may be part of this by being involved in defining the structure information.

Nucleolar–associating domains (*NADs*) may help determine structure. They are primarily genomic regions with heterochromatic signatures and include transposons, sub–telomeric regions and mostly inactive protein–coding genes (Pontvianne et al., 2016). There are two classes of NADs: NAD–1 is associated with nuclear lamina and is involved in gene repression (Vertii et al., 2018). This type of NAD may be related to structure determination.

The CCCTC–binding factor (*CTCF*) is used in chromatin remodeling and the positioning of nucleosomes. This binding of this transcription factor has been shown to be affected by methylation of CpG islands (Teif et al., 2014). This could be involved in the process of structure determination since CTCF binding is shown to be specific to cell type. CTCF binding sites are often found in SINEs (Schmidt et al., 2012) (Chuong et al., 2017) indicating that CTCF is used in structure determination.

CTCF is involved in controlling the boundaries of active and inactive chromatin in the Hox clusters, affecting the expression of genes (Narendra et al., 2015). It has been shown that TADs are bound by CTCF barrier elements. It has been noted that SINE element retrotransposition may alter these CTCF binding sites (Dixon et al., 2012). Cell type specificity may arise from rearrangements in local chromatin structure that allow for different patterns and insulation capacity of long–range CTCF interactions (Narendra et al., 2016).

The mechanism for controlling cell typing and structural determination involves the Polycomb Repressive Complex (*PRC*) genes and the Trithorax Group (*TrxG*) (Khalil et al., 2009). They affect the chromatin and histone to select the correct cell type. PRC2 seems to turn off unneeded structures. This is especially true for the gross structure determination of the HOX gene related transposons (Rinn et al., 2007) (Walter, 2015). TrxG proteins work to activate gene expression related to non–coding RNA (Sha and Boyer, 2009) (Maeda and Karch, 2009).

Here are some examples of the effects of transposons and chromatin remodeling. One study has shown that a SINE B2 transposon is implicated in the expression of growth hormone in the developing pituitary gland (Lunyak et al., 2007) by imposing a local perturbation in chromatin structure resulting in repositioning of the gene from a heterochromatic region to a more permissive euchromatic region. Another study shows that a SINE regulates a growth factor in the brain (Okada et al., 2010). In a third study, errors in the development of alveoli where shown to be caused by alterations in a lnc–RNA LINC01081 which controlled the expression of the FOXF1 transcription enhancer (Szafranski et al., 2014). Finally,

the analysis of lnc–RNA taurine upregulated gene 1 (Tug1) revealed approximately 400 genes that were positively regulated and approximately 560 genes that were negatively regulated by Tug1. Errors in this regulation can result in diabetic nephropathy (Long et al., 2016).

Processes such as regeneration are shown to be controlled by histone modification. Studies of leg regeneration in Gryllus bimaculatus shows that changing the H3K27 methylation state starts regeneration. This regeneration is position–specific. If the leg amputation position is changed, the regeneration is different (Hamada et al., 2015).

It is important to note that if this hypothesis is correct, the process of structure determination is completely different from the process of gene activation, which also involves the manipulation of chromatin and histone modification. Since the two processes involve two different encodings and are for the most part located in two different parts of the genome (genetic versus intergenetic regions) they probably do not interfere with each other if gene expression occurs in the euchromatin and structure determination occurs in the heterochromatin. In support of this, it has been observed that a much larger proportion of histone–mark–defined enhancers overlap transposons than gene expression enhancers (Simonti et al., 2017).

Note that three–dimensonal changes in a TAD turn on or off families of genes. This is different from the three–dimensional changes for structure determination. Instead of trying to find the cell type, this manipulation of the TAD is to turn on the correct genes for that TAD once the cell type has been determined.

Since heterochromatin is not typically involved in the transcription of genes, it does not need to be lightly packed like the euchromatin. All that is required is that it facilitates the determination of a single cell type encoding datum which is implemented through the histone. The tight packing of heterochromatin could actually serve its purpose by making it better able to model the three–dimensional structure.

This is a complex process, with different types of transposons and lnc–RNA doing their part. For example, LINE–1 may be a subroutine that is part of the larger structure (Brouha et al., 2003). LINE–1 is known to be active in neuronal differentiation (Faulkner and Billon, 2018). Also, the regulatory pathways could be activated directly by transposons that contain binding sites for some transcription factors besides being activated indirectly by histone modification (Bourque et al., 2008). The activation of transcription factors is specific to the individual transposons, indicating that these factors are essential for certain cell types (Sundaram et al., 2014).

As the structure is being developed, it is also possible for the modifications made by the transposons to be undone. It is thought that the insertion and removal of transposons are imprecise, which affects surrounding DNA sequences (Bourque et al., 2008). These insertions could actually be very precise. They could be part of the process by which the specific cell identification is made as the context shifts. Erasures could probably be associated with a switch of local context as the cell determines its precise location in the structure. This could result in histone de–methylation (Sha and Boyer, 2009).

The Histone Code hypothesis (Strahl and Allis, 2000) (Jenuwein and Allis, 2001) is that specific patterns of modifications to the histone tail are read like a molecular bar code to recruit the cellular machinery that brings about a distinct chromatin state (Cosgrove and Wolberger, 2005) (Prakash and Fournier, 2017). Once the final location of a cell in a structure is determined and thus the type of cell it is, these alterations may be used to initiate the regulation of the proteins associated with that particular cell type. This is especially true for the H3 Histone. Transposon–derived active enhancers are enriched for a suite of individual histone modifications – H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K27ac (Huda et al., 2011) (Cao et al., 2019).

The process may work as follows. It has been observed that unique chromatin signatures correlate with cell type and function (Sha and Boyer, 2009). Studies have shown that there are histone modifications even in intergenic regions (Rosenfeld et al., 2009). Methylation of DNA represses structure determination, which affects whether the histone is modified (Bannister and Kouzarides, 2011). Modification of the DNA sequence by transposons identifies the cell type in the structure, making this information available to the histone. Histone modification is involved in implementing the cell type identification (Sha and Boyer, 2009) by regulating the transcription of the genes related to that cell type.

Part of the process of structure development is apoptosis. This helps to complete the boundaries of a structure and its inner voids. It has been noted that LINE–1 elements are associated with apoptosis (Kapusta et al., 2013). This indicates that apoptosis is an integral part of structure determination.

Reprogramming (dedifferentiation) could be considered a process of turning off structure information.

This is accomplished by DNA demethylation (Reik et al., 2001) (Miyoshi et al., 2016).

IMPLICATIONS OF STRUCTURE ENCODING

Determination of cell fate in a structure is a significant factor in embryogenesis. As the embryo develops, cells proceed from stem cells to pluripotent, then multipotent cells and finally to unipotent cells. The previous sections lay out a hypothesis on how the steps in structure definition contribute to this progression. Now we will consider some implications of this hypothesis.

Structure encoding is associated with the musculoskeletal system, the construction of organs and the neural structures in the brain. Obviously, structure information is highly tissue–specific. This hypothesis states that the expression of transposons and lnc–RNAs is, therefore, tissue–specific (Cabili et al., 2011) (Yan et al., 2013) (Liu et al., 2016). For example, lnc–RNAs have been studied in the regulation of mammary gland development and endocrine signalling (Sun and Kraus, 2013).

The brain and nervous system have a structure that is far too complex to be defined just by the use of proteins — even in complex combinations. There are many special purpose regions, each with a fine structure that differs from other areas.

Instincts are also coded for. They tend to be very specific, detailed behaviors that can vary even within a species, if the species is separated into isolated groups. Instincts are obviously passed on as traits common to all members of the local population and could easily be specified in the brain structure, neuron by neuron. This is an essential feature of Sociobiology (Wilson, 2000).

Within a species, there are usually significant differences in structure between the sexes. Therefore, methylation, both at CpG sites and with the histones, are involved in genomic imprinting (Pask et al., 2009) (Ferguson-Smith, 2011) (Ikeda and Nishimura, 2015). Transposons and non–coding RNAs are activated differently for the sexes, which results in differential cell typing and the expression of the associated genes (McDonald et al., 2005) (Autuoro et al., 2014). The imprinting control regions are non–coding, consistent with the structure hypothesis. If they are deleted the imprinted genes are not expressed (Mancini-DiNardo et al., 2006) (Bartolomei, 2009).

Since structure information is not needed in the germline, transposons are normally turned off (Walter, 2015) (Haig, 2016). As the embryo starts to develop, they are turned on (Gerdes et al., 2016).

There are believed to be over 200 different types of somatic cells in the human body (Patel and Yang, 2010). The encoding of body parts does not have to specify all of these unique cell type values, though. A particular organ, such as a lung, kidney, muscle, or skeletal structure may have its own encoding which would specify one of a family of cell types that have a particular set of active genes in common, Fine structure determination of cell type would then result in a fully differentiated cell.

Systems such as the circulatory or nervous system may have structure information, but only for the larger parts of the system such as the heart or the brain. The determination of the finer structure should include the insights of *Facilitated Variation* (Gerhart and Kirschner, 2007). The theory of facilitated variation claims that the evolution of anatomical and physiological traits are the result of regulatory changes in the usage of various members of a large set of conserved core components that function in development and physiology. This theory points out that the structure information does not have to specify every detail. For example, the circulatory system is constructed at the detailed level in an ad hoc manner, as the body part develops. For example, it is known that there is significant variation between people in the anatomy of the arteries that supply the heart (Ogobuiro and Tuma, 2018). Facilitated variation also points out the advantage of modular design. With properly designed modularity, variation within each module can be generated without harming other modules (Parter et al., 2008).

The hypothesis of structure encoding extends Facilitated Variation theory. The structure of the major systems encode for the cell types that are specific to the given system, but ancillary support systems are added as needed. The structures are organized hierarchically, which leads to modular design.

There are a variety of diseases associated with the activity of transposons and non–coding RNA (Cordaux and Batzer, 2009). These diseases often happen because the transposon has caused an insertion in a working gene, which is not its purpose according to this hypothesis. The purpose of a transposon is to determine structure — any direct changes to genes by transposons could probably be an error.

Some, maybe even most, cancers could be due to errors in structure definition. This could include errors in the process of methylization (Lim and Maher, 2010), transposon splicing, or chromatin modification (Sun et al., 2018). Also, it has been noted that LINE–1 transposons may be involved in the initiation of cancer (Scott et al., 2016) and the progression of cancer that has already started (Hancks and Kazazian,

2016). Errors in communicating structure information may also cause cancer. Abnormal exosomal long intergenic non–protein–coding RNAs have been implicated in cancer, indicating communication problems (Dragomir et al., 2018).

EVOLUTION AND SPECIATION

Transposons and lnc–RNA represent a major source of lineage–specific DNA and thus is a significant factor in speciation (Kapusta et al., 2013). This is due to the evolution of body structure leading to diversification.

A study of *Drosophila melanogaster* showed that adaptive changes to transposons for populations in Africa versus North America showed predominately changes in introns or intergenic regions. There was no clear overriding pattern in the types of genes that are located near the adaptive transposons (González et al., 2008). If the transposons are encoding for structure, then it would not be likely that there is a difference in the expression of proteins — rather, there would be a difference in morphology.

It has been observed that non–coding RNA is changing at a faster rate than protein–coding regions (Kutter et al., 2012) (Kapusta and Feschotte, 2014) (Chen et al., 2016). Lnc–RNA is shown to have tissue specificity, being different for colon, spleen, lung, testes, brain, kidney, liver, heart, and skeletal muscle, regardless of the species in which they were profiled. A significant number of lnc–RNAs are conserved across mammals, but there are as many as 20% that are unique to humans and possibly chimpanzee (Washietl et al., 2014). It has also been noted that the evolutionary trajectory of transposons in mammals is similar across species despite clade–specific differences (Buckley et al., 2017).

Studies of Human Accelerated Regions (*HARs*) show that most HARs lie in the intergenic and intronic non–coding regions (Doan et al., 2016). 97% of HARs are noncoding — 92% are found in intergenic regions and introns (Levchenko et al., 2017). Some are lnc–RNAs. Nearly half of all HARs function as enhancers of neural progenitor cells (Ryu et al., 2018). Of 510 regions conserved between chimpanzee and macaque, but deleted in humans, almost all reside in noncoding regions.

Studies have shown that the density of transposons located in the Hox clusters leads to higher speciation rates (Feiner, 2016). Also, the number of transposons associated with Hox clusters varies between different phyla of vertebrates (Di-Poï et al., 2009).

Since transposons play an important part in embryonic development by implementing structural determination, they are evolutionarily conserved, compared to other intergenic regions (Kapusta et al., 2013).

It has been noted that the genomic ratio between SINEs and LINE–2 are highly correlated across different mammalian genomes (Cao et al., 2019).

Some non–coding RNA is highly conserved (Bejerano et al., 2004) (Katzman et al., 2007). This implies that these mutations are deleterious and thus do not spread much in populations. This indicates that they play an essential part in the function of the organism — possibly because these sequences are structures that are essential to any organism. Analysis shows that conserved non–coding elements appear more often in the intergenic regions. They are also known to include transposons (Makunin et al., 2013). This is not true of all structures, though: mutations in the structure information is probably a major driving force in the evolution of animals (Bourque et al., 2008).

VERIFICATION AND EXPERIMENTAL TESTS

Structural determination is not part of the continuous processing in the cell. These determinations need only be made in the initial ontogenesis. Consequently, it is possible that transposons would not be active very much relative to the regular processes in the cell, but more so in its formation (Kapranov and St Laurent, 2012) (Palazzo and Lee, 2015).

For organisms that have pairs of genes, it is likely that only one of the copies of the structure in each chromosome pair is used. This would be because, instead of the problem of integrating two differing sequences into a unified structure, only one sequence from one chromosome is chosen and the other discarded. Therefore, detailed structure information may not have the dominant and recessive characteristics of genes. Note that not all sequences from one chromosome need to be chosen — this decision could be made on a case–by–case basis.

Many inherited diseases may not be due to mutations in genes but in structures. It is quite likely that diseases that are believed to be due to multiple genetic errors may, in fact, be due to a single structural

error.

This may be why schizophrenia is passed in twins only about 50% of the time, because this disorder may come from structural problems, instead of being a problem with the proteins (Gejman et al., 2010). In that case, the disorder is manifested if the malformed structure is chosen.

Another lnc–RNA associated with a disease state is myocardial infarction associated transcript (*MIAT*). Increased expression of MIAT is related to severe inflammatory dilated chronic cardiomyopathy from Chagas disease (Frade et al., 2016). Although this may be due to vascular dysfunction, it may have a structural cause.

The *Caenorhabditis elegans* genome contains both transposons and non–coding RNA. As much as 12% of the *C. elegans* genome are transposons (Bessereau, 2006). There are approximately 1,300 non–coding RNAs also (Stricklin et al., 2005) of which about 170 are lnc–RNAs (Nam and Bartel, 2012). Studies have shown that some of these RNA sequences have physiological functioning including chromatin modification (Akay et al., 2019). They also show different expression patterns for different cell types (Liu et al., 2017). The presence or absence of a transposon in *C. elegans* is found to be associated with phenotypic differences (Laricchia et al., 2017). These results are consistent with the claims of this hypothesis.

It is possible to develop a mapping between transposons and their associated non–coding RNA and the body part they define by looking at organisms that are missing that part or where it is malformed.

CONCLUSIONS

This hypothesis suggests that the cellular DNA of plants and animals contains both procedural and declarative information, each with their own individual encoding. In evolutionary terms, primordial cell DNA contained mostly procedural information. As multicellular organisms formed and developed, declarative information was added to specify structure.

This hypothesis forces a reinterpretation of how transposons and non–coding RNA function. Instead of directly controlling expression, they affect the chromatin, which alters the histone. The Histone Code defines a cell type which then results in the gene expression for that type of cell.

The hypothesis explains why transposons are prevalent in the genome. Instead of being considered unwanted causes of genetic disease, they actually play an essential part in the embryogenesis of plants and animals. Also, this hypothesis revises the distinction between euchromatin versus heterochromatin. Whereas euchromatin is where the expression of individual genes occur, heterochromatin is designed to make it possible to determine the specific location of an individual cell in a body part.

It is difficult to interpret many genetic studies in relation to this hypothesis because the underlying assumptions of the studies do not match the hypothesis. That is to say, the Central Dogma that the DNA directly controls the expression of RNA which determines the proteins that get expressed means that if elements like transposons and non–coding RNA are assumed to be involved in controlling the expression of genes, then by the dogma they are regulatory elements that directly control expression. Instead, this hypothesis states that they are indicators of cell types, which then result in the regulation of genes according to that type via the action of the histone. If correct, that means that some of the processes that result in cell typing are misinterpreted. For example, sometimes what is considered DNA breakage repair may actually be part of the process of three–dimensional structure determination by transposon splicing (Felsenfeld, 2014). Also, studies of gene enhancers need to be interpreted in light of this hypothesis. Certain genes are enhanced because the particular cell type at that particular location requires it.

Bacterial transposons may possibly be a precursor to the transposons in multicellular organisms. Evolution exapted this mechanism as a way of constructing complex multicellular organisms on top of the gross structure of morphogens.

The Cambrian explosion (Gould, 1990) could be explained by the evolution of a structure mechanism. The gross structures appeared before this time, but the creation of the ability to define a fine structure could have led to an ability to fine–tune the organism to a particular environmental niche.

REFERENCES

Akay, A., Jordan, D., Navarro, I. C., Wrzesinski, T., Ponting, C. P., Miska, E. A., and Haerty, W. (2019). Identification of functional long non-coding RNAs in C. elegans. *BMC biology*, 17(1):14.

Almouzni, G. and Cedar, H. (2016). Maintenance of epigenetic information. *Cold Spring Harbor Perspectives in Biology*, 8(5):a019372.

Autuoro, J., Pirnie, S., and Carmichael, G. (2014). Long noncoding RNAs in imprinting and X chromosome inactivation. *Biomolecules*, 4(1):76–100.

Babenko, V. N., Chadaeva, I. V., and Orlov, Y. L. (2017). Genomic landscape of cpg rich elements in human. *BMC evolutionary biology*, 17(1):19.

Bannister, A. J. and Kouzarides, T. (2011). Regulation of chromatin by histone modifications. *Cell research*, 21(3):381.

Bartolomei, M. S. (2009). Genomic imprinting: employing and avoiding epigenetic processes. *Genes & development*, 23(18):2124–2133.

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W. J., Mattick, J. S., and Haussler, D. (2004). Ultraconserved elements in the human genome. *Science*, 304(5675):1321–1325.

Bessereau, J.-L. (2006). Transposons in C. elegans. *WormBook*, 18:1–13.

Bourque, G., Burns, K. H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H. L., Macfarlan, T. S., et al. (2018). Ten things you should know about transposable elements. *Genome biology*, 19(1):199.

Bourque, G., Leong, B., Vega, V. B., Chen, X., Lee, Y. L., Srinivasan, K. G., Chew, J.-L., Ruan, Y., Wei, C.-L., Ng, H. H., et al. (2008). Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome research*, 18(11):1752–1762.

Brosius, J. (1991). Retroposons—seeds of evolution. *Science*, 251(4995):753.

Brosius, J. and Gould, S. J. (1992). On" genomenclature": a comprehensive (and respectful) taxonomy for pseudogenes and other" junk DNA". *Proceedings of the National Academy of Sciences*, 89(22):10706–10710.

Brouha, B., Schustak, J., Badge, R. M., Lutz-Prigge, S., Farley, A. H., Moran, J. V., and Kazazian, H. H. (2003). Hot L1s account for the bulk of retrotransposition in the human population. *Proceedings of the National Academy of Sciences*, 100(9):5280–5285.

Buckley, R. M., Kortschak, R. D., Raison, J. M., and Adelson, D. L. (2017). Similar evolutionary trajectories for retrotransposon accumulation in mammals. *Genome biology and evolution*, 9(9):2336–2353.

Buldyrev, S., Goldberger, A., Havlin, S., Mantegna, R., Matsa, M., Peng, C.-K., Simons, M., and Stanley, H. (1995). Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis. *Physical Review E*, 51(5):5084.

Bürglin, T. R. and Affolter, M. (2016). Homeodomain proteins: an update. *Chromosoma*, 125(3):497–521.

Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J. L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & development*, 25(18):1915–1927.

Cao, Y., Chen, G., Wu, G., Zhang, X., McDermott, J., Chen, X., Xu, C., Jiang, Q., Chen, Z., Zeng, Y., Ai, D., Huang, Y., and Han, J.-D. J. (2019). Widespread roles of enhancer-like transposable elements in cell identity and long-range genomic interactions. *Genome Research*, 29(1):40–52.

Chen, Hongjunand Du, G., Song, X., and Li, L. (2017). Non-coding transcripts from enhancers: new insights into enhancer activity and gene expression regulation. *Genomics, Proteomics and Bioinformatics*, 15(3):201–207.

Chen, J., Shishkin, A. A., Zhu, X., Kadri, S., Maza, I., Guttman, M., Hanna, J. H., Regev, A., and Garber, M. (2016). Evolutionary analysis across mammals reveals distinct classes of long non-coding rnas. *Genome biology*, 17(1):19.

Chuong, E. B., Elde, N. C., and Cédric, F. (2017). Regulatory activities of transposable elements: from conflicts to benefits. *Nature Reviews Genetics*, 18(2):71–86.

Cordaux, R. and Batzer, M. A. (2009). The impact of retrotransposons on human genome evolution. *Nature Reviews Genetics*, 10(10):691.

Cosgrove, M. S. and Wolberger, C. (2005). How does the histone code work? *Biochemistry and Cell Biology*, 83(4):468–476.

Cournac, A., Koszul, R., and Mozziconacci, J. (2015). The 3D folding of metazoan genomes correlates with the association of similar repetitive elements. *Nucleic acids research*, 44(1):245–255.

Crick, F. (1970). Central dogma of molecular biology. *Nature*, 227(5258):561.

De Laat, W. and Duboule, D. (2013). Topology of mammalian developmental enhancers and their

regulatory landscapes. *Nature*, 502(7472):499.
Dekker, J. and Heard, E. (2015). Structural and functional diversity of topologically associating domains. *FEBS letters*, 589(20PartA):2877–2884.
Di-Poï, N., Montoya-Burgos, J. I., and Duboule, D. (2009). Atypical relaxation of structural constraints in Hox gene clusters of the green anole lizard. *Genome research*, 19(4):602–610.
Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376.
Doan, R. N., Bae, B.-I., Cubelos, B., Chang, C., Hossain, A. A., Al-Saad, S., Mukaddes, N. M., Oner, O., Al-Saffar, M., Balkhy, S., Gascon, G. G., Nieto, M., and Walsh, C. A. (2016). Mutations in human accelerated regions disrupt cognition and social behavior. *Cell*, 167(2):341–354.
Dragomir, M., Chen, B., and Calin, G. A. (2018). Exosomal lncRNAs as new players in cell-to-cell communication. *Translational cancer research*, 7(Suppl 2):S243.
Durruthy-Durruthy, J., Sebastiano, V., Wossidlo, M., Cepeda, D., Cui, J., Grow, E. J., Davila, J., Mall, M., Wong, W. H., Wysocka, J., et al. (2016). The primate-specific noncoding RNA HPAT5 regulates pluripotency during human preimplantation development and nuclear reprogramming. *Nature genetics*, 48(1):44.
Ezkurdia, I., Juan, D., Rodriguez, J. M., Frankish, A., Diekhans, M., Harrow, J., Vazquez, J., Valencia, A., and Tress, M. L. (2014). Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. *Human molecular genetics*, 23(22):5866–5878.
Faulkner, G. J. and Billon, V. (2018). L1 retrotransposition in the soma: a field jumping ahead. *Mobile DNA*, 9(1):22.
Feiner, N. (2016). Accumulation of transposable elements in Hox gene clusters during adaptive radiation of Anolis lizards. *Proceedings of the Royal Society B: Biological Sciences*, 283(1840):20161555.
Felsenfeld, G. (2014). A brief history of epigenetics. *Cold Spring Harbor perspectives in biology*, 6(1):a018200.
Felsenfeld, G. and Groudine, M. (2003). Controlling the double helix. *Nature*, 421(6921):448.
Ferguson-Smith, A. C. (2011). Genomic imprinting: the emergence of an epigenetic paradigm. *Nature Reviews Genetics*, 12(8):565.
Feschotte C, P. E. (2007). DNA transposons and the evolution of eukaryotic genomes. *Annual Review of Genetics*, 41:331–368.
Fortin, J.-P. and Hansen, K. D. (2015). Reconstructing a/b compartments as revealed by hi-c using long-range correlations in epigenetic data. *Genome biology*, 16(1):180.
Frade, A. F., Laugier, L., Ferreira, L. R. P., Baron, M. A., Benvenuti, L. A., Teixeira, P. C., Navarro, I. C., Cabantous, S., Ferreira, F. M., da Silva Cândido, D., et al. (2016). Myocardial infarction–associated transcript, a long noncoding RNA, is overexpressed during dilated cardiomyopathy due to chronic chagas disease. *The Journal of infectious diseases*, 214(1):161–165.
Gejman, P. V., Sanders, A. R., and Duan, J. (2010). The role of genetics in the etiology of schizophrenia. *Psychiatric Clinics*, 33(1):35–66.
Gerdes, P., Richardson, S. R., Mager, D. L., and Faulkner, G. J. (2016). Transposable elements in the mammalian embryo: pioneers surviving through stealth and service. *Genome biology*, 17(1):100.
Gerhart, J. and Kirschner, M. (2007). The theory of facilitated variation. *Proceedings of the National Academy of Sciences*, 104(suppl 1):8582–8589.
González, J., Lenkov, K., Lipatov, M., Macpherson, J. M., and Petrov, D. A. (2008). High rate of recent transposable element–induced adaptation in drosophila melanogaster. *PLoS biology*, 6(10):e251.
Gould, S. J. (1990). *Wonderful life: the Burgess Shale and the nature of history*. WW Norton & Company.
Gray, H. (1977). *Gray's Anatomy. The Classic Collector's Edition*. New York, Bounty Books.
Haig, D. (2016). Transposable elements: Self-seekers of the germline, team-players of the soma. *BioEssays*, 38(11):1158–1166.
Hamada, Y., Bando, T., Nakamura, T., Ishimaru, Y., Mito, T., Noji, S., Tomioka, K., and Ohuchi, H. (2015). Leg regeneration is epigenetically regulated by histone H3K27 methylation in the cricket Gryllus bimaculatus. *Development*, 142(17):2916–2927.
Hancks, D. C. and Kazazian, H. H. (2016). Roles for retrotransposon insertions in human disease. *Mobile DNA*, 7(1):9.
Huda, A., Tyagi, E., Mariño-Ramírez, L., Bowen, N., Jjingo, D., and Jordan, I. (2011). Prediction of

transposable element derived enhancers using chromatin modification profiles. *PloS one*, 6(11):e27513.

Iimura, T. and Pourquié, O. (2007). Hox genes in time and space during vertebrate body formation. *Development, growth & differentiation*, 49(4):265–275.

Ikeda, Y. and Nishimura, T. (2015). The role of DNA methylation in transposable element silencing and genomic imprinting. In *Nuclear functions in plant transcription, signaling and development*, pages 13–29. Springer.

Jacques, P.-E., Jeyakani, J., and Bourque, G. (2013). The majority of primate-specific regulatory sequences are derived from transposable elements. *PLoS genetics*, 9(5):e1003504.

Jenuwein, T. and Allis, C. D. (2001). Translating the histone code. *Science*, 293(5532):1074–1080.

Jin, B., Li, Y., and Robertson, K. D. (2011). DNA methylation: superior or subordinate in the epigenetic hierarchy? *Genes & cancer*, 2(6):607–617.

Kapranov, P. and St Laurent, G. (2012). Dark matter RNA: existence, function, and controversy. *Frontiers in genetics*, 3:60.

Kapusta, A. and Feschotte, C. (2014). Volatile evolution of long noncoding rna repertoires: mechanisms and biological implications. *Trends in Genetics*, 30(10):439–452.

Kapusta, A., Kronenberg, Z., Lynch, V. J., Zhuo, X., Ramsay, L., Bourque, G., Yandell, M., and Feschotte, C. (2013). Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS genetics*, 9(4):e1003470.

Katan-Khaykovich, Y. and Struhl, K. (2005). Heterochromatin formation involves changes in histone modifications over multiple cell generations. *The EMBO journal*, 24(12):2138–2149.

Katzman, S., Kern, A. D., Bejerano, G., Fewell, G., Fulton, L., Wilson, R. K., Salama, S. R., and Haussler, D. (2007). Human genome ultraconserved elements are ultraselected. *Science*, 317(5840):915–915.

Kazazian, H. H. (2011). Mobile DNA transposition in somatic cells. *BMC biology*, 9(1):62.

Kerszberg, M. and Wolpert, L. (2007). Specifying positional information in the embryo: looking beyond morphogens. *Cell*, 130(2):205–209.

Khalil, A. M., Guttman, M., Huarte, M., Garber, M., Raj, A., Morales, D. R., Thomas, K., Presser, A., Bernstein, B. E., van Oudenaarden, A., et al. (2009). Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proceedings of the National Academy of Sciences*, 106(28):11667–11672.

Kobayashi, H., Lowe, M., and Kikyo, N. (2017). Epigenetic regulation of open chromatin in pluripotent stem cells. In *Translating Epigenetics to the Clinic*, pages 1–18. Elsevier.

Kragesteen, B. K., Spielmann, M., Paliou, C., Heinrich, V., Schöpflin, R., Esposito, A., Annunziatella, C., Bianco, S., Chiariello, A. M., Jerković, I., et al. (2018). Dynamic 3d chromatin architecture contributes to enhancer specificity and limb morphogenesis. *Nature genetics*, 50(10):1463.

Kutter, C., Watt, S., Stefflova, K., Wilson, M. D., Goncalves, A., Ponting, C. P., Odom, D. T., and Marques, A. C. (2012). Rapid turnover of long noncoding rnas and the evolution of gene expression. *PLoS genetics*, 8(7):e1002841.

Laricchia, K., Zdraljevic, S., Cook, D., and Andersen, E. C. (2017). Natural variation in the distribution and abundance of transposable elements across the Caenorhabditis elegans species. *Molecular biology and evolution*, 34(9):2187–2202.

Lawrence, P. A. and Struhl, G. (1996). Morphogens, compartments, and pattern: lessons from drosophila? *Cell*, 85(7):951–961.

Lempradl, A. and Ringrose, L. (2008). How does noncoding transcription regulate hox genes? *Bioessays*, 30(2):110–121.

Levchenko, A., Kanapin, A., Samsonova, A., and Gainetdinov, R. R. (2017). Human accelerated regions and other human-specific sequence variations in the context of evolution and their relevance for brain development. *Genome biology and evolution* , 10(1):166–188.

Lieberman-Aiden, E., Van Berkum, N., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B., Sabo, P., Dorschner, M., and Sandstrom, R. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293.

Lim, D. H. and Maher, E. R. (2010). DNA methylation: a form of epigenetic control of gene expression. *The Obstetrician & Gynaecologist*, 12(1):37–42.

Liu, S. J., Nowakowski, T. J., Pollen, A. A., Lui, J. H., Horlbeck, M. A., Attenello, F. J., He, D., Weissman, J. S., Kriegstein, A. R., Diaz, A. A., et al. (2016). Single-cell analysis of long non-coding RNAs in the developing human neocortex. *Genome biology*, 17(1):67.

Liu, W., Yu, E., Chen, S., Ma, X., Lu, Y., and Liu, X. (2017). Spatiotemporal expression profiling of long intervening noncoding RNAs in Caenorhabditis elegans. *Scientific reports*, 7(1):5195.

Long, J., Badal, S. S., Ye, Z., Wang, Y., Ayanga, B. A., Galvan, D. L., Green, N. H., Chang, B. H., Overbeek, P. A., and Danesh, F. R. (2016). Long noncoding rna tug1 regulates mitochondrial bioenergetics in diabetic nephropathy. *The Journal of clinical investigation*, 126(11):4205–4218.

Lunyak, V. V., Prefontaine, G. G., Núñez, E., Cramer, T., Ju, B.-G., Ohgi, K. A., Hutt, K., Roy, R., García-Díaz, A., Zhu, X., et al. (2007). Developmentally regulated activation of a SINE B2 repeat as a domain boundary in organogenesis. *Science*, 317(5835):248–251.

Maeda, R. K. and Karch, F. (2009). The bithorax complex of drosophila: an exceptional hox cluster. *Current topics in developmental biology*, 88:1–33.

Makunin, I. V., Shloma, V. V., Stephen, S. J., Pheasant, M., and Belyakin, S. N. (2013). Comparison of ultra-conserved elements in drosophilids and vertebrates. *PloS one*, 8(12):e82362.

Mancini-DiNardo, D., Steele, S. J., Levorse, J. M., Ingram, R. S., and Tilghman, S. M. (2006). Elongation of the Kcnq1ot1 transcript is required for genomic imprinting of neighboring genes. *Genes & development*, 20(10):1268–1282.

Mateo, L. J., Murphy, S. E., Hafner, A., Cinquini, I. S., Walker, C. A., and Boettiger, A. N. (2019). Visualizing dna folding and rna in embryos at single-cell resolution. *Nature*, page 1.

Mattick, J. S. (2003). Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. *Bioessays*, 25(10):930–939.

Mattick, J. S. (2009). The genetic signatures of noncoding RNAs. *PLoS genetics*, 5(4):e1000459.

McCue, A. D. and Slotkin, R. K. (2012). Transposable element small rnas as regulators of gene expression. *Trends in genetics*, 28(12):616–623.

McDonald, J., Matzke, M., and Matzke, A. (2005). Host defenses to transposable elements and the evolution of genomic imprinting. *Cytogenetic and genome research*, 110(1–4):242–249.

Messerschmidt, D. M., Knowles, B. B., and Solter, D. (2014). DNA methylation dynamics during epigenetic reprogramming in the germline and preimplantation embryos. *Genes & development*, 28(8):812–828.

Miyoshi, N., Stel, J. M., Shioda, K., Qu, N., Odajima, J., Mitsunaga, S., Zhang, X., Nagano, M., Hochedlinger, K., Isselbacher, K. J., et al. (2016). Erasure of DNA methylation, genomic imprints, and epimutations in a primordial germ-cell model derived from mouse pluripotent stem cells. *Proceedings of the National Academy of Sciences*, 113(34):9545–9550.

Nam, J.-W. and Bartel, D. P. (2012). Long noncoding RNAs in C. elegans. *Genome research*, 22(12):2529–2540.

Narendra, V., Bulajić, M., Dekker, J., Mazzoni, E. O., and Reinberg, D. (2016). Ctcf-mediated topological boundaries during development foster appropriate gene regulation. *Genes & development*, 30(24):2657–2662.

Narendra, V., Rocha, P. P., An, D., Raviram, R., Skok, J. A., Mazzoni, E. O., and Reinberg, D. (2015). Ctcf establishes discrete functional chromatin domains at the hox clusters during differentiation. *Science*, 347(6225):1017–1021.

Negre, B. and Simpson, P. (2013). Diversity of transposable elements and repeats in a 600 kb region of the fly Calliphora vicina. *Mobile DNA*, 4(1):13.

Noordergraaf, L. and Boldyshev, K. (2013). Hello, world! http://asm.sourceforge.net/intro/hello.html. Accessed: February 24, 2019.

Ogobuiro, I. and Tuma, F. (2018). Anatomy, Thorax, Heart Coronary Arteries. In *StatPearls [Internet]*. StatPearls Publishing.

Okada, N., Sasaki, T., Shimogori, T., and Nishihara, H. (2010). Emergence of mammals by emergency: exaptation. *Genes to Cells*, 15(8):801–812.

Ørom, U. A. and Shiekhattar, R. (2011). Noncoding RNAs and enhancers: complications of a long-distance relationship. *Trends in Genetics*, 27(10):433–439.

Palazzo, A. F. and Lee, E. S. (2015). Non-coding RNA: what is functional and what is junk? *Frontiers in genetics*, 6:2.

Parter, M., Kashtan, N., and Alon, U. (2008). Facilitated variation: how evolution learns from past environments to generalize to new environments. *PLoS computational biology*, 4(11):e1000206.

Pask, A. J., Papenfuss, A. T., Ager, E. I., McColl, K. A., Speed, T. P., and Renfree, M. B. (2009). Analysis of the platypus genome suggests a transposon origin for mammalian imprinting. *Genome biology*,

10(1):R1.

Patel, M. and Yang, S. (2010). Advances in reprogramming somatic cells to induced pluripotent stem cells. *Stem Cell Reviews and Reports*, 6(3):367–380.

Peric-Hupkes, D., Meuleman, W., Pagie, L., Bruggeman, S. W., Solovei, I., Brugman, W., Gräf, S., Flicek, P., Kerkhoven, R. M., van Lohuizen, M., et al. (2010). Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. *Molecular cell*, 38(4):603–613.

Piriyapongsa, J., Mariño-Ramírez, L., and Jordan, I. K. (2007). Origin and evolution of human microRNAs from transposable elements. *Genetics*, 176(2):1323–1337.

Pontvianne, F., Carpentier, M.-C., Durut, N., Pavlištová, V., Jaške, K., Schořová, Š., Parrinello, H., Rohmer, M., Pikaard, C. S., Fojtová, M., et al. (2016). Identification of nucleolus-associated chromatin domains reveals a role for the nucleolus in 3D organization of the A. thaliana genome. *Cell reports*, 16(6):1574–1587.

Pope, B. D., Ryba, T., Dileep, V., Yue, F., Wu, W., Denas, O., Vera, D. L., Wang, Y., Hansen, R. S., Canfield, T. K., et al. (2014). Topologically associating domains are stable units of replication-timing regulation. *Nature*, 515(7527):402.

Prakash, K. and Fournier, D. (2017). Deciphering the histone code to build the genome structure. *Preprint biorXiv:10.1101/217190v2*.

Rebollo, R., Karimi, M. M., Bilenky, M., Gagnier, L., Miceli-Royer, K., Zhang, Y., Goyal, P., Keane, T. M., Jones, S., Hirst, M., et al. (2011). Retrotransposon-induced heterochromatin spreading in the mouse revealed by insertional polymorphisms. *PLoS genetics*, 7(9):e1002301.

Reik, W., Dean, W., and Walter, J. (2001). Epigenetic reprogramming in mammalian development. *Science*, 293(5532):1089–1093.

Rinn, J. L., Kertesz, M., Wang, J. K., Squazzo, S. L., Xu, X., Brugmann, S. A., Goodnough, L. H., Helms, J. A., Farnham, P. J., Segal, E., et al. (2007). Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell*, 129(7):1311–1323.

Rodriguez-Campos, A. and Azorín, F. (2007). RNA is an integral component of chromatin that contributes to its structural organization. *PloS one*, 2(11):e1182.

Rosenfeld, J. A., Wang, Z., Schones, D. E., Zhao, K., DeSalle, R., and Zhang, M. Q. (2009). Determination of enriched histone modifications in non-genic portions of the human genome. *BMC genomics*, 10(1):143.

Ryu, H., Inoue, F., Whalen, S., Williams, A., Kircher, M., Martin, B., Alvarado, B., Samee, M. A. H., Keough, K., Thomas, S., Kriegstein, A., Shendure, J., Pollen, A., Ahituv, N., and Pollard, K. S. (2018). Massively parallel dissection of human accelerated regions in human and chimpanzee neural progenitors. *Preprint biorXiv:10.1101/256313*.

Sarge, K. D. and Park-Sarge, O.-K. (2005). Gene bookmarking: keeping the pages open. *Trends in biochemical sciences*, 30(11):605–610.

Schmidt, D., Schwalie, P. C., Wilson, M. D., Ballester, B., Gonçalves, Â., Kutter, C., Brown, G. D., Marshall, A., Flicek, P., and Odom, D. T. (2012). Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell*, 148(1–2):335–348.

Scott, E. C., Gardner, E. J., Masood, A., Chuang, N. T., Vertino, P. M., and Devine, S. E. (2016). A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer. *Genome research*, 26(6):745–755.

Sha, K. and Boyer, L. A. (2009). The chromatin signature of pluripotent cells. *Stem Book. Cambridge: Harvard Stem Cell Institute*.

Shevelyov, Y. Y. and Nurminsky, D. I. (2012). The nuclear lamina as a gene-silencing hub. *Current issues in molecular biology*, 14(1):27.

Simonti, C. N., Pavličev, M., and Capra, J. A. (2017). Transposable element exaptation into regulatory regions is rare, influenced by evolutionary age, and subject to pleiotropic constraints. *Molecular biology and evolution*, 34(11):2856–2869.

Strahl, B. D. and Allis, C. D. (2000). The language of covalent histone modifications. *Nature*, 403(6765):41.

Stricklin, S. L., Griffiths-Jones, S., and Eddy, S. R. (2005). C. elegans noncoding RNA genes. *WormBook*, 25:1–7.

Sun, M. and Kraus, W. L. (2013). Minireview: Long noncoding rnas: new "links" between gene expression and cellular outcomes in endocrinology. *Molecular endocrinology*, 27(9):1390–1402.

Sun, X., Wang, X., Tang, Z., Grivainis, M., Kahler, D., Yun, C., Mita, P., Fenyö, D., and Boeke, J. D. (2018). Transcription factor profiling reveals molecular choreography and key regulators of human retrotransposon expression. *Proceedings of the National Academy of Sciences*, 115(24):E5526–E5535.

Sundaram, V., Cheng, Y., Ma, Z., Li, D., Xing, X., Edge, P., Snyder, M. P., and Wang, T. (2014). Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome research*, 24(12):1963–1976.

Szafranski, P., Dharmadhikari, A. V., Wambach, J. A., Towe, C. T., White, F. V., Grady, R. M., Eghtesady, P., Cole, F. S., Deutsch, G., Sen, P., et al. (2014). Two deletions overlapping a distant foxf1 enhancer unravel the role of lncrna linc01081 in etiology of alveolar capillary dysplasia with misalignment of pulmonary veins. *American Journal of Medical Genetics Part A*, 164(8):2013–2019.

Teif, V. B., Beshnova, D. A., Vainshtein, Y., Marth, C., Mallm, J.-P., Höfer, T., and Rippe, K. (2014). Nucleosome repositioning links DNA (de) methylation and differential CTCF binding during stem cell development. *Genome research*, 24(8):1285–1295.

Teif, V. B., Vainshtein, Y., Caudron-Herger, M., Mallm, J.-P., Marth, C., Höfer, T., and Rippe, K. (2012). Genome-wide nucleosome positioning during embryonic stem cell development. *Nature structural & molecular biology*, 19(11):1185.

Tetta, C., Ghigo, E., Silengo, L., Deregibus, M. C., and Camussi, G. (2013). Extracellular vesicles as an emerging mechanism of cell-to-cell communication. *Endocrine*, 44(1):11–19.

Thompson, P. J., Macfarlan, T. S., and Lorincz, M. C. (2016). Long terminal repeats: from parasitic elements to building blocks of the transcriptional regulatory repertoire. *Molecular cell*, 62(5):766–776.

Tropberger, P. and Schneider, R. (2013). Scratching the (lateral) surface of chromatin regulation by histone modifications. *Nature structural & molecular biology*, 20(6):657.

Tsumagari, K., Baribault, C., Terragni, J., Chandra, S., Renshaw, C., Sun, Z., Song, L., Crawford, G. E., Pradhan, S., Lacey, M., et al. (2013). DNA methylation and differentiation: HOX genes in muscle cells. *Epigenetics & chromatin*, 6(1):25.

Van Steensel, B. and Belmont, A. S. (2017). Lamina-associated domains: links with chromosome architecture, heterochromatin, and gene repression. *Cell*, 169(5):780–791.

Vertii, A., Ou, J., Yu, J., Yan, A., Liu, H., Zhu, L. J., Kaufman, P. D., et al. (2018). Two Contrasting Classes of Nucleolus-Associated Domains in Mouse Fibroblast Heterochromatin. *Preprint bioRxiv:10.1101/484568v1*.

Walter, M. (2015). *Transposon regulation upon dynamic loss of DNA methylation*. PhD thesis, Université Pierre et Marie Curie-Paris VI.

Wardle, F. C. and Smith, J. C. (2004). Refinement of gene expression patterns in the early Xenopus embryo. *Development*, 131(19):4687–4696.

Washietl, S., Kellis, M., and Garber, M. (2014). Evolutionary dynamics and tissue specificity of human long noncoding rnas in six mammals. *Genome research*, 24(4):616–628.

Wikipedia contributors (2018a). Central dogma of molecular biology. [Online; accessed 13-February-2019].

Wikipedia contributors (2018b). Gene. [Online; accessed 28-December-2018].

Wilson, E. O. (2000). *Sociobiology: The New Synthesis*. Harvard University Press.

Winograd, T. (1975). Frame representations and the procedural/declarative controversy. *Representation and understanding: studies in cognitive science*, pages 185–210.

Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., Liu, P., Lian, Y., Zheng, X., Yan, J., et al. (2013). Single-cell rna-seq profiling of human preimplantation embryos and embryonic stem cells. *Nature structural & molecular biology*, 20(9):1131.

Yang, Y., Wen, L., and Zhu, H. (2015). Unveiling the hidden function of long non-coding RNA by identifying its major partner-protein. *Cell & bioscience*, 5(1):59.

Zemach, A., McDaniel, I. E., Silva, P., and Zilberman, D. (2010). Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science*, 328(5980):916–919.