



## A Novel Approach to Genome Editing using Cellular Automata Evolutions of Adjoints Sequences

By Prashanthi Govindarajan, Sathya Govindarajan & Ethirajan Govindarajan

*Abstract-* This paper proposes a novel method for genome editing using cellular automata evolutions of adjoints of Adenine, Thymine, Guanine, and Cytosine. The adjoints of the given a genome sequence are the characteristic binary string sequences. For example, the adjoint of Adenine of a given genome sequence is a binary string consisting of 0's and 1's where 1's corresponds to the presence of Adenine in the genome sequence. So, one can have four adjoint sequences of Adenine, Thymine, Guanine, and Cytosine corresponding to a given genome sequence. One-dimensional three neighborhood binary value cellular automata rules can be applied to an adjoint sequence and the desired number of evolutions could be obtained. This rule is defined by a linear Boolean function and one can have 256 such linear Boolean functions. Genome editing is carried out by superimposing the evolved adjoint sequence on the original genome sequence or on its successive evolutions. In this manner, one can have four ways of genome editing using four adjoint sequences and evolutions.

*Keywords:* genome editing, cellular automata, evolutions of adjoints, linear boolean functions.

*GJCST-G Classification:* J.3



*Strictly as per the compliance and regulations of:*



# A Novel Approach to Genome Editing using Cellular Automata Evolutions of Adjoint Sequences

Prashanthi Govindarajan<sup>α</sup>, Sathya Govindarajan<sup>σ</sup> & Ethirajan Govindarajan<sup>ρ</sup>

**Abstract-** This paper proposes a novel method for genome editing using cellular automata evolutions of adjoints of Adenine, Thymine, Guanine, and Cytosine. The adjoints of the given a genome sequence are the characteristic binary string sequences. For example, the adjoint of Adenine of a given genome sequence is a binary string consisting of 0's and 1's where 1's corresponds to the presence of Adenine in the genome sequence. So, one can have four adjoint sequences of Adenine, Thymine, Guanine, and Cytosine corresponding to a given genome sequence. One-dimensional three neighborhood binary value cellular automata rules can be applied to an adjoint sequence and the desired number of evolutions could be obtained. This rule is defined by a linear Boolean function and one can have 256 such linear Boolean functions. Genome editing is carried out by superimposing the evolved adjoint sequence on the original genome sequence or on its successive evolutions. In this manner, one can have four ways of genome editing using four adjoint sequences and evolutions.

**Keywords:** genome editing, cellular automata, evolutions of adjoints, linear boolean functions.

## I. INTRODUCTION

Genome editing is essentially the process of introducing required changes in a given DNA. A protein or enzyme cuts certain portions of a given DNA and substitutes the target sequences of nucleotides by specific chains of nucleotides. For ages, people have been working on the notion of Genome editing. However, only recently, the spurt of activities has been reported in the literature about genome editing. Before getting into the details of the genome editing, it would be apt to reason out why genome editing is viewed as a significant activity. Genome editing is carried out as a health initiative to heal even the so-called incurable diseases associated with genetic problems. Clinically there are many methods of editing genomes among which CRISPR technique is recognized as a reliable technique ratified by the U.S. Food and Drug Administration, pushing in a new era of cancer treatment. CRISPR based therapy is designed to treat blood and bone marrow cancer, which usually affects children and young adults. This therapy is known

as CAR-T therapy, and it has shown remarkable results in patients. By eliminating those genes which cause disease, physicians can treat various illnesses ranging from heart disease to Alzheimer's.

In addition to curing diseases, gene therapy (gene editing) could be used to stop inherited disease in its tracks to save endangered species, and more so, to resurrect extinct species. All such gene therapy techniques are clinical laboratory- based. Alternatively, one can try out the possibilities of developing some of them for genome editing using computational tools and concepts. Genome is a string of nucleotides, and its characteristic sequence is a sequence of A, T, G, and C. Gene editing, in the computer science point of view, is a pattern searching and substituting process, meaning, genome editing is essentially a string processing operation. A genome is a subset of a free monoid  $A^*$  of an alphabet  $A$  which consists of the primitive symbols A, T, G and C. So, gene editing is viewed as a map  $\phi$  that connects  $A^*$  to  $A^*$ . It is in this context, this paper introduces a novel concept of genome editing cellular automata evolutions of adjoint strings of a genome. Section 2 of this paper describes the fundamental notions of adjoints of a genome and their evolution using one-dimensional cellular automata rules defined by linear Boolean functions. Section 3 presents the technique of editing genome sequences using the cellular automata generations of adjoint sequences. Section 4 illustrates the concept with the help of a case study.

## II. ADJOINTS OF A GENOME SEQUENCE AND THEIR EVOLUTIONS USING ONE-DIMENSIONAL THREE NEIGHBORHOOD CELLULAR AUTOMATA

Adjoint of a particular nucleotide in a genome sequence is the binary sequence obtained by substituting the particular nucleotides in the genome sequence by 1's and the others by 0's. For example, let us consider a sample sequence of BrucellaSuis 1330 for a case study. The actual length of the genome sequence of BrucellaSuis 1330 is 5806. A cellular automaton is an idealized parallel processing system consisting of an array of numbers (1-D, 2-D and more) realized using updating rules based on certain

Author <sup>α</sup>: Chief Executive Officer, PRC Global Technologies Inc., Ontario, Canada. e-mail: sathyag413@gmail.com

Author <sup>σ ρ</sup>: Director, PRC Global Technologies Inc., Ontario, Canada. e-mails: prashanthi89@gmail.com, dr.rajaneg@gmail.com

neighborhood. For example, a one dimensional cellular automaton would consist of a finite length array as shown below.



Consider an *i*th cell in the array. This cell has a neighbor *i*-1 on its left and another *i*+1 on its right. All three put together is called a three neighborhood. One can assign a site (cell) variable  $\xi_{i-1}$ ,  $\xi_i$ , and  $\xi_{i+1}$  to the three neighborhood cells. At a particular instant of time, these variables take on numerical values, say either a 0 or a 1. In such a case, the variables are denoted as  $\xi_{ti-1}$ ,  $\xi_{ti}$ , and  $\xi_{ti+1}$ . The value of the *i*th cell at the next instant of time is evaluated using an updating rule that involves the present values of the *i*th, (*i*-1)th and (*i*+1)th cells. This updating rule is basically a linear Boolean function of three variables. One can construct 256 linear Boolean functions, as updating rules of one-dimensional three-neighborhood binary-valued cellular automata. Each rule defines an automaton by itself. So, one dimensional binary-valued three-neighborhood cellular automata (123CA) rules could be used to model adjoints of a genome sequence. The first twenty linear Boolean functions of cellular automata 123CA are listed below with their decimal equivalents.

Linear Boolean Function	Decimal Equivalent
0	0
$(\bar{\xi}_{i-1}\bar{\xi}_i\bar{\xi}_{i+1})$	1
$(\bar{\xi}_{i-1}\bar{\xi}_i\xi_{i+1})$	2
$(\bar{\xi}_{i-1}\xi_i)$	3
$(\bar{\xi}_{i-1}\xi_i\bar{\xi}_{i+1})$	4
$(\bar{\xi}_{i-1}\xi_{i+1})$	5
$(\bar{\xi}_{i-1}\xi_i\xi_{i+1}) + (\bar{\xi}_{i-1}\bar{\xi}_i\xi_{i+1})$	6
$(\bar{\xi}_{i-1}\xi_{i+1}) + (\bar{\xi}_{i-1}\bar{\xi}_i)$	7
$(\bar{\xi}_{i-1}\xi_i\xi_{i+1})$	8
$(\bar{\xi}_{i-1}\bar{\xi}_i\xi_{i+1}) + (\bar{\xi}_{i-1}\xi_i\xi_{i+1})$	9
$(\bar{\xi}_{i-1}\xi_{i+1})$	10
$(\bar{\xi}_{i-1}\xi_i) + (\bar{\xi}_{i-1}\xi_{i+1})$	11
$(\bar{\xi}_{i-1}\xi_i)$	12
$(\bar{\xi}_{i-1}\xi_{i+1}) + (\bar{\xi}_{i-1}\xi_i)$	13
$(\bar{\xi}_{i-1}\xi_i) + (\bar{\xi}_{i-1}\xi_{i+1})$	14
$(\bar{\xi}_{i-1})$	15
$(\bar{\xi}_{i-1}\bar{\xi}_i\xi_{i+1})$	16
$(\bar{\xi}_i\xi_{i+1})$	17
$(\bar{\xi}_{i-1}\bar{\xi}_i\xi_{i+1}) + (\bar{\xi}_{i-1}\bar{\xi}_i\xi_{i+1})$	18
$(\bar{\xi}_i\xi_{i+1}) + (\bar{\xi}_{i-1}\bar{\xi}_i)$	19
$(\bar{\xi}_{i-1}\bar{\xi}_i\xi_{i+1}) + (\bar{\xi}_{i-1}\bar{\xi}_i\xi_{i+1})$	20

For the case study rule number 90 is applied to the adjoints of BrucellaSuis 1330 genome sequence and 500 evolutions generated. Rule 90 is shown below.

$$(\bar{\xi}_{i-1}\bar{\xi}_{i+1}) + (\bar{\xi}_{i-1}\xi_{i+1}) \quad 90$$

Since the image of the 500 evolutions of BrucellaSuis 1330 is quite large, a small portion of the

images are presented in this paper. Fig. 1 shows evolutions of the adjoints of A(n) and T(n).

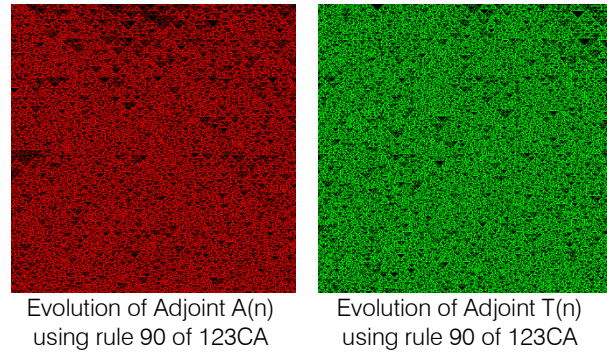


Fig. 1: Evolutions of A(n) and T(n)

Fig. 2 shows evolution of the adjoints of G(n) and C(n).

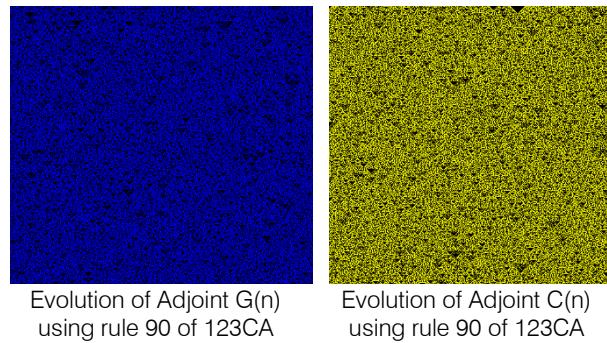


Fig. 2: Evolutions of G (n) and C (n)

### III. NUCLEOTIDE ADJOINTS BASED GENOME EDITING

As outlined earlier, a genome sequence is a subset of a free monoid  $A^*$  of an alphabet  $A$  which consists of the primitive symbols A, T, G and C and gene editing is a map  $\phi$  that connects  $A^*$  to  $A^*$ . The map  $\phi$  is a rule that transforms a sequence into another desired sequence. One can use four types of symbol substitution formulas given below for genome editing.

Symbol to be substituted	→	Substitution symbol	Formula Number
$A \vee T \vee G \vee C$	→	A	1
$A \vee T \vee G \vee C$	→	T	2
$A \vee T \vee G \vee C$	→	G	3
$A \vee T \vee G \vee C$	→	C	4

Where the symbol  $\vee$  denotes the relation of logical OR. Application of formula #1 to any genome sequence is called A-latch. Similarly one can think of T-latch, G-latch and C-latch. A nucleotide latch would give rise to conversion of any genome sequence into a sequence consisting of that particular nucleotide in one step. This is called 'Nucleotide Saturation'. Thus one can have A-saturation, T-saturation, G-saturation, and C-saturation. To be precise, A-latching of any genome sequence transforms it into A-saturated sequence and it

holds for other three types of latching also. The question that arises here is that whether it is possible to edit a genome sequence using nucleotide latching techniques at preferred locations in the sequence. One such possibility is latching of a genome sequence using cellular automata evolutions of adjoints. Once a genome sequence is latched, the resulting edited sequence is called its elementary transformation. Two types of latching are discussed here (i) latching of elementary transformations of a genome sequence with various cellular automata evolutions of adjoints and (ii) latching of original genome sequence with various cellular automata evolutions of adjoints. Fig. 3 portrays the first approach of latching of elementary transformations of a genome sequence with various cellular automata evolutions of adjoints.

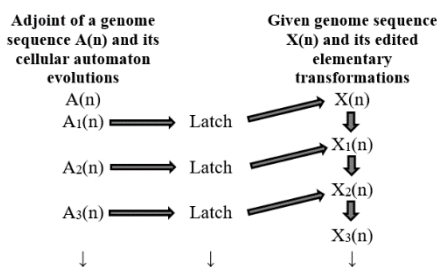


Fig. 3: Editing of elementary transformations of a genome sequence

Fig. 4 portrays the second approach of latching of original genome sequence with various cellular automata evolutions of adjoints.

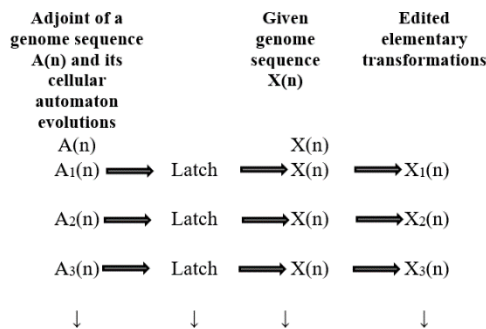


Fig. 4: Editing of original genome sequence

#### IV. CASE STUDY

The characteristic sequence of BrucellaSuis 1330 genome sequence is used here for the case study. The length of this sequence is 5806. Rule number 90 is used here for the study. The linear Boolean function corresponding to this rule is

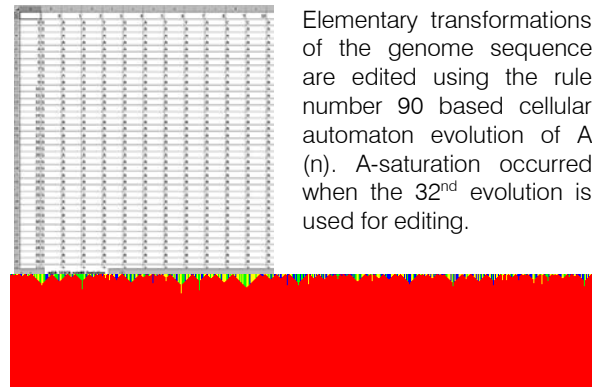
$$(\xi_{i-1}\bar{\xi}_{i+1}) + (\bar{\xi}_{i-1}\xi_{i+1})$$

##### Approach #1

Editing of elementary transformations of a genome sequence with various cellular automata evolutions of adjoints

Figs. 5 to 8 portrait the first approach of latching of elementary transformations of a genome sequence with various cellular automata evolutions of adjoints.

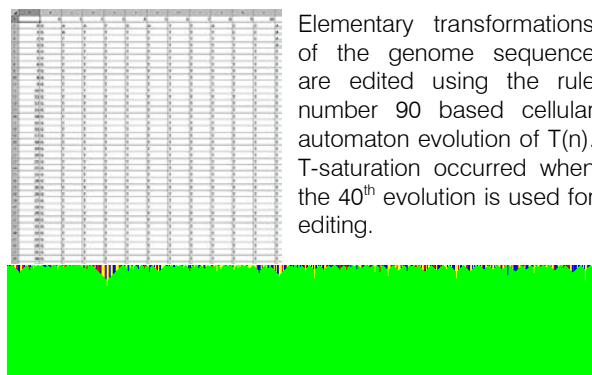
Fig. 5 shows the result of A-latching the genome sequence, that is, Adenine based genome editing. Since the image is quite large, a small portion of it is shown here. It was observed that the A-saturation of the genome sequence occurred while editing the previous elementary transformation of the genome sequence using the 32nd evolution of the A(n).



Elementary transformations of the genome sequence are edited using the rule number 90 based cellular automaton evolution of A (n). A-saturation occurred when the 32<sup>nd</sup> evolution is used for editing.

Fig. 5: Adenine based genome editing (A-saturation at 32<sup>nd</sup> evolution)

Fig. 6 shows the result of T-latching the genome sequence, that is, Thymine based genome editing. Since the image is quite large, a small portion of it is shown here. It was observed that the T-saturation of the genome sequence occurred while editing the previous elementary transformation of the genome sequence using the 40<sup>th</sup> evolution of the A(n).

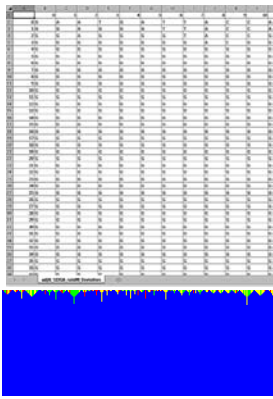


Elementary transformations of the genome sequence are edited using the rule number 90 based cellular automaton evolution of T(n). T-saturation occurred when the 40<sup>th</sup> evolution is used for editing.

Fig. 6: Thymine based genome editing (T-saturation at 40<sup>th</sup> evolution)

Fig. 7 shows the result of G-latching the genome sequence, that is, Guanine based genome editing. Since the image is quite large, a small portion of it is shown here. It was observed that the G-saturation of the genome sequence occurred while editing the previous elementary transformation of the genome sequence using the 15<sup>th</sup> evolution of the G(n).

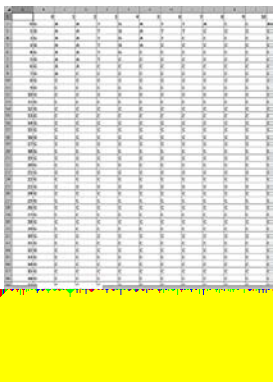




Elementary transformations of the genome sequence are edited using the rule number 90 based cellular automaton evolution of GT(n). G-saturation occurred when the 15<sup>th</sup> evolution is used for editing.

Fig. 7: Guanine based genome editing (G-saturation at 15<sup>th</sup> evolution)

Fig. 8 shows the result of C-latching the genome sequence, that is, Cytosine based genome editing. Since the image is quite large, a small portion of it is shown here. It was observed that the C-saturation of the genome sequence occurred while editing the previous elementary transformation of the genome sequence using the 21<sup>st</sup> evolution of the C(n).



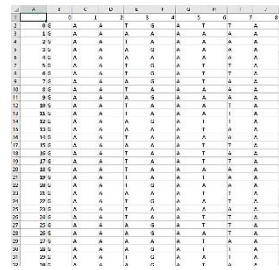
Elementary transformations of the genome sequence are edited using the rule number 90 based cellular automaton evolution of C(n). C-saturation occurred when the 21<sup>st</sup> evolution is used for editing.

Fig. 8: Cytosine based genome editing (C-saturation at 21<sup>st</sup> evolution)

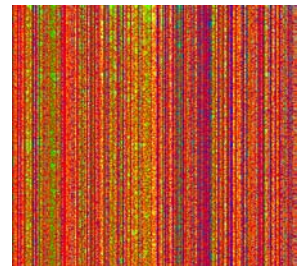
**Approach #2**

Editing of a genome sequence with various cellular automata evolutions of adjoints

Figs. 9 to 14 portrait the second approach of latching of a given genome sequence with rule number 90 based cellular automata evolutions of adjoints. Figs. 9 shows Adenine based genome editing on original sequence both in Text Form and in Image Form. The Image Form is obtained using a color coding scheme which paints a red color for Adenine, green color for Thymine, blue color for Guanine and yellow color for Cytosine. Fig. 10 shows the predominant horizontal lines and vertical lines of the image of the edited genome separately. An image processing tool of line detector is used for this purpose.

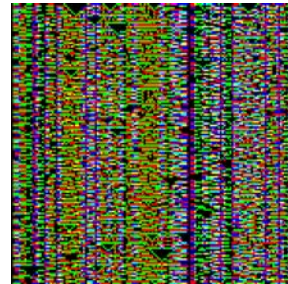


Adenine based genome editing on original sequence (Text Form)

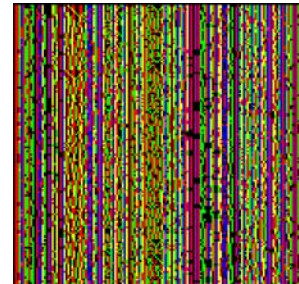


Adenine based genome editing on original sequence (Image Form)

Fig. 9: Edited genome using evolutions of A(n)



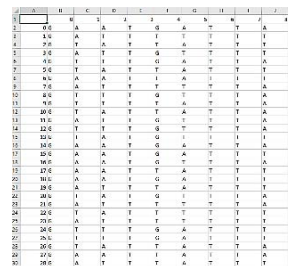
Adenine based genome editing on original sequence Horizontal lines Detected



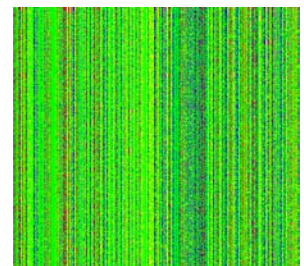
Adenine based genome editing on original sequence Vertical lines Detected

Fig. 10: Horizontal and Vertical lines of A-latch detected

Figs. 11 shows Thymine based genome editing on original sequence both in Text Form and in Image Form. Fig. 12 shows the predominant horizontal lines and vertical lines of the image of the edited genome separately.

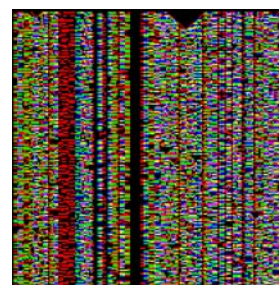


Thymine based genome editing on original sequence (Text Form)

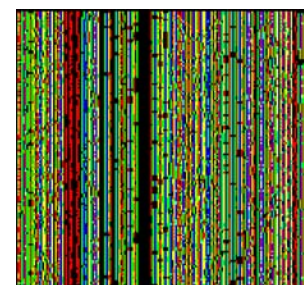


Thymine based genome editing on original sequence (Image Form)

Fig. 11: Edited genome using evolutions of T(n)



Thymine based genome editing on original sequence Horizontal lines Detected

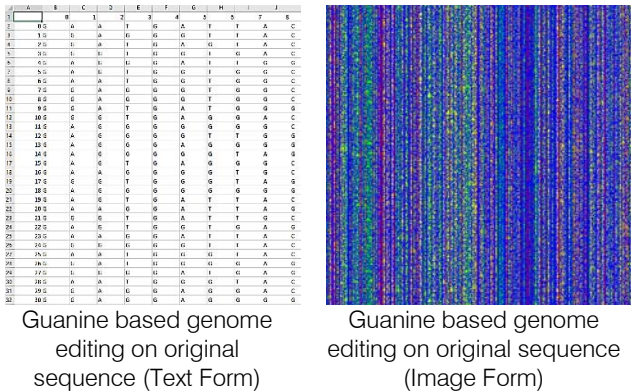


Thymine based genome editing on original sequence Vertical lines Detected

Fig. 12: Horizontal and Vertical lines of T-latch detected

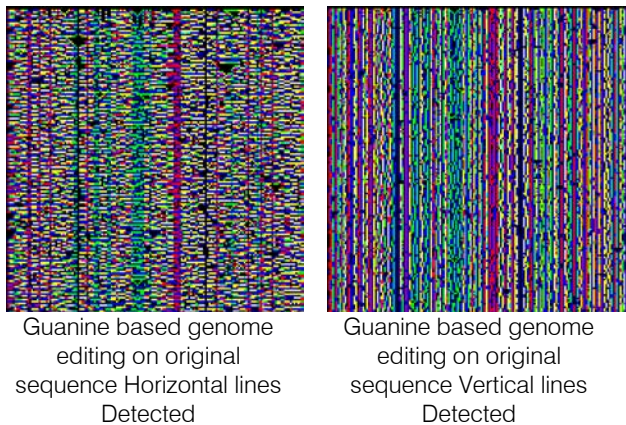


Figs. 13 shows Guanine based genome editing on original sequence both in Text Form and in Image Form. Fig. 14 shows the predominant horizontal lines and vertical lines of the image of the edited genome separately.



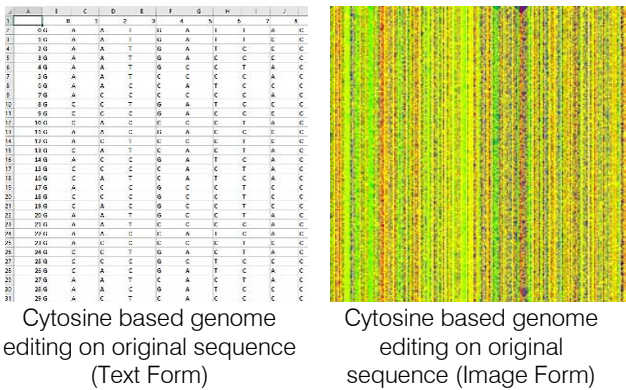
Guanine based genome editing on original sequence (Text Form)      Guanine based genome editing on original sequence (Image Form)

Fig.13: Edited genome using evolutions of G(n)



Guanine based genome editing on original sequence Horizontal lines Detected      Guanine based genome editing on original sequence Vertical lines Detected

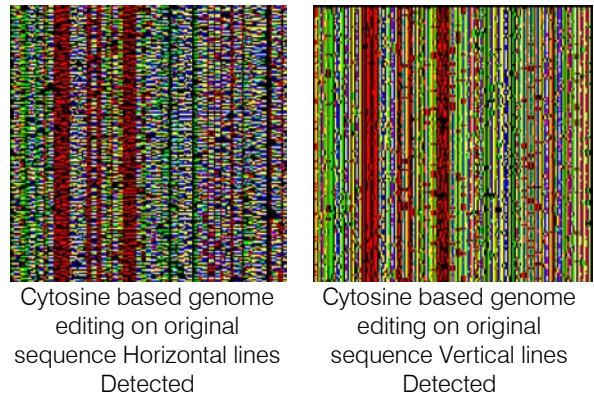
Fig.14: Horizontal and Vertical lines of G-latch detected



Cytosine based genome editing on original sequence (Text Form)      Cytosine based genome editing on original sequence (Image Form)

Fig.15: Edited genome using evolutions of C(n)

Figs. 15 shows Cytosine based genome editing on original sequence both in Text Form and in Image Form. Fig. 16 shows the predominant horizontal lines and vertical lines of the image of the edited genome separately.



Cytosine based genome editing on original sequence Horizontal lines Detected      Cytosine based genome editing on original sequence Vertical lines Detected

Fig.16: Horizontal and Vertical lines of C-latch detected

Dyadic Operations among edited genomes

Two relational operations 'max' and 'min' are carried out on the A-latch, T-latch, G-latch and C-latch edited genomes. Adenine is represented by the number 1, Thymine by 2, Guanine by 3 and Cytosine by 4. Now, the relation max(x,y) of two nucleotides is evaluated as the maximum of the numerical values of the nucleotides under comparison. Similarly, the relation min(x,y) of two nucleotides is evaluated as the minimum of the numerical values of the nucleotides under comparison. Fig. 17 shows the result of comparing A-latch and T-latch of BrucellaSuis 1330 genome with the max operator along with its horizontal and vertical lines.

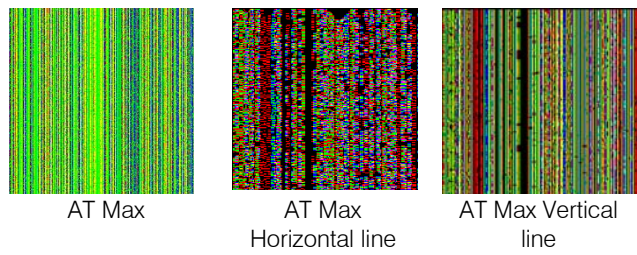


Fig.17: Max of A and T latches and its horizontal and vertical lines

Fig. 18 shows the result of comparing A-latch and C-latch of BrucellaSuis 1330 genome with the max operator along with its horizontal and vertical lines.

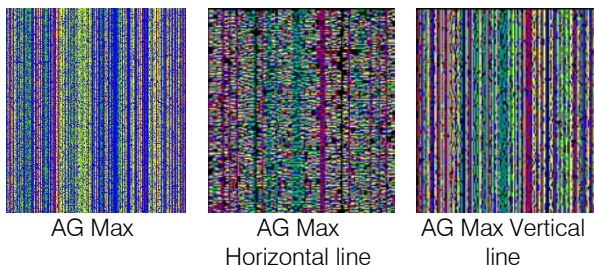
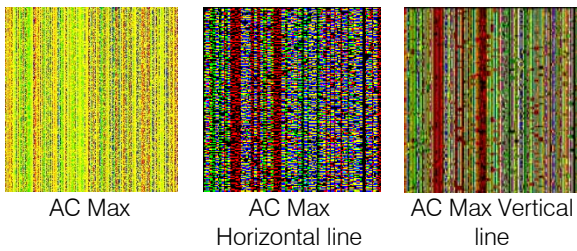


Fig.18: Max of A and G latches and its horizontal and vertical lines

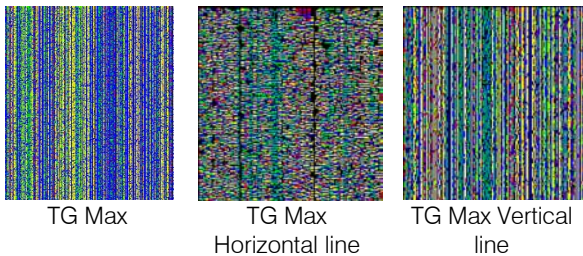
Fig. 19 shows the result of comparing A-latch and C-latch of BrucellaSuis 1330 genome with the max operator along with its horizontal and vertical lines.





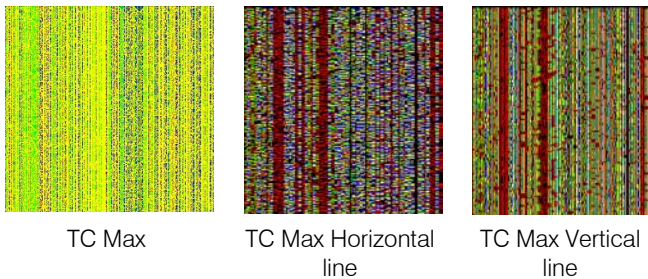
*Fig. 19:* Max of A and C latches and its horizontal and vertical lines

Fig. 20 shows the result of comparing T-latch and G-latch of BrucellaSuis 1330 genome with the max operator along with its horizontal and vertical lines.



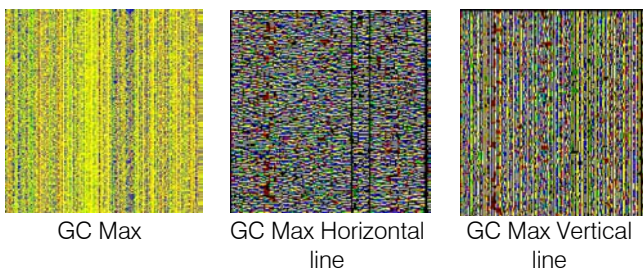
*Fig. 20:* Max of T and G latches and its horizontal and vertical lines

Fig. 21 shows the result of comparing T-latch and C-latch of BrucellaSuis 1330 genome with the max operator along with its horizontal and vertical lines.



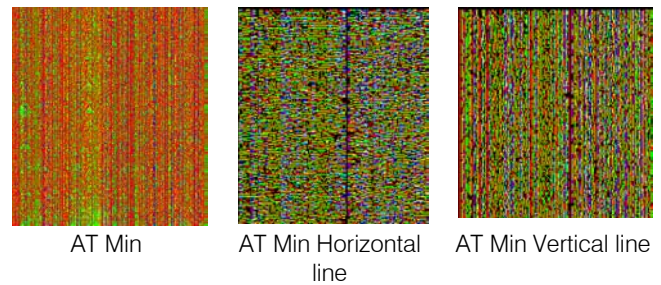
*Fig. 21:* Max of T and C latches and its horizontal and vertical lines

Fig. 22 shows the result of comparing G-latch and C-latch of BrucellaSuis 1330 genome with the max operator along with its horizontal and vertical lines.



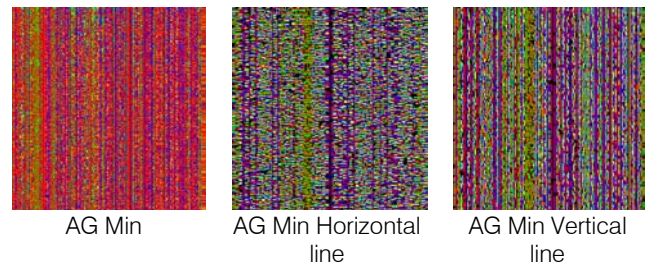
*Fig. 22:* Max of G and C latches and its horizontal and vertical lines

Fig. 23 shows the result of comparing A-latch and T-latch of BrucellaSuis 1330 genome with the min operator along with its horizontal and vertical lines.



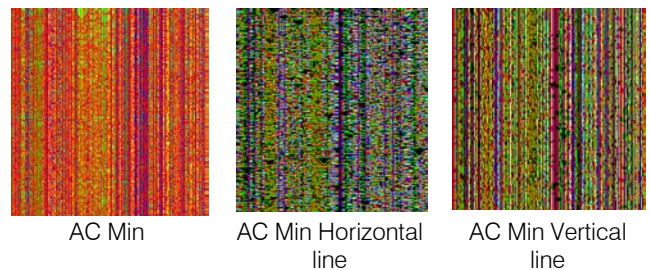
*Fig. 23:* Min of A and T latches and its horizontal and vertical lines

Fig. 24 shows the result of comparing A-latch and G-latch of BrucellaSuis 1330 genome with the min operator along with its horizontal and vertical lines.



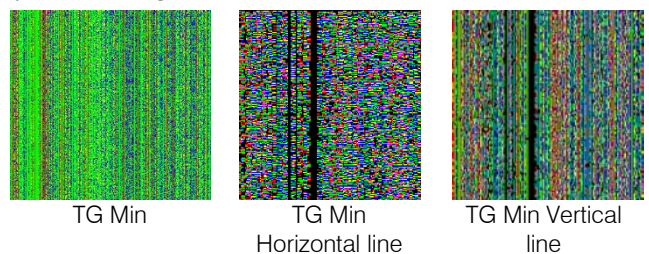
*Fig. 24:* Min of A and G latches and its horizontal and vertical lines

Fig. 25 shows the result of comparing A-latch and C-latch of BrucellaSuis 1330 genome with the min operator along with its horizontal and vertical lines.



*Fig. 25:* Min of A and C latches and its horizontal and vertical lines

Fig. 26 shows the result of comparing T-latch and G-latch of BrucellaSuis 1330 genome with the min operator along with its horizontal and vertical lines.



*Fig. 26:* Min of T and G latches and its horizontal and vertical lines

Fig. 27 shows the result of comparing T-latch and C-latch of *BrucellaSuis* 1330 genome with the min operator along with its horizontal and vertical lines. Fig. 28 shows the result of comparing G-latch and C-latch of *BrucellaSuis* 1330 genome with the min operator along with its horizontal and vertical lines.

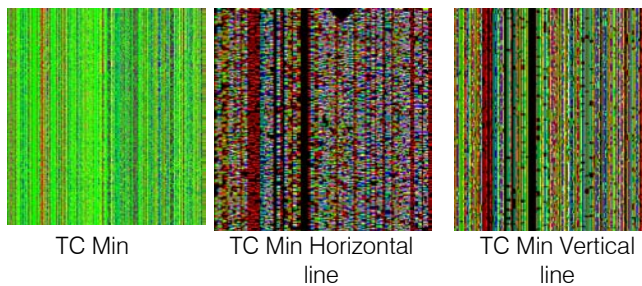


Fig. 27: Min of T and C latches and its horizontal and vertical lines

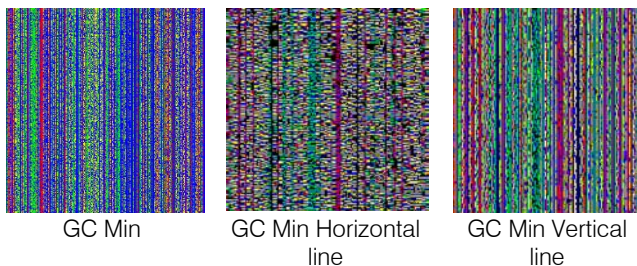


Fig. 28: Min of G and C latches and its horizontal and vertical lines

## V. RESULTS AND DISCUSSIONS

The following observations are made from the case study.

- Editing a genome is a map or a function that transforms a genome into a desired genome sequence
- Two approaches could be undertaken for genome editing (i) Editing of elementary transformations of a genome sequence with various cellular automata evolutions of adjoints and(ii) Editing of the given genome sequence with various cellular automata evolutions of adjoints
- Using approach 1, one would end up with saturation in short steps. Using approach 2, one would be able to generate edited versions that exhibit periodicities and generic evolutions.
- Relational operations like 'max' and 'min' could be carried out on various latches. One can have 11 such max operations like (i) A-latch and T-latch, (ii) A-latch and G-latch, (iii) A-latch and C-latch, (iv) T-latch and G-latch, (v) T-latch and C-latch, (vi) G-latch and C-latch, (vii) A-latch, T-latch and G-latch (viii) A-latch, T-latch and C-latch(ix) A-latch, G-latch and C-latch, (x)T-latch, G-latch and C-latch, (xi) A-latch,T-latch, G-latch and C-latch. In the same manner, one can have 11 min operations.
- The result of each operation exhibits unique behavior in periodicity and generic form.
- Interpretation of these properties could be made in a better way only by an expert in genetic Engineering

## ACKNOWLEDGEMENTS

The authors express their profound gratitude to the management of Pentagram Research Centre Private Limited, Hyderabad for their boundless support in carrying out research in their premises. The authors further put on record the invaluable guidance of Prof. Dr E G Rajan, President of Pentagram Research and data scientists Mr. Rahul Sharma, Mr. Srikanth Maddikunta and Mr. Shubham Karande.

## REFERENCES RÉFÉRENCES REFERENCIAS

- "Worlds Record Breaking Plant: Deletes its Noncoding "Junk" DNA". Design & Trend. May 12, 2013. Retrieved 2013-06-04.
- Pennisi, E. (6 September 2012). "ENCODE Project Writes Eulogy for Junk DNA". *Science* 337 (6099): 1159–1161. doi:10.1126/science.337.6099.1159. PMID 22955811.
- The ENCODE Project Consortium (2012). "An integrated encyclopedia of DNA elements in the human genome". *Nature* 489 (7414): 57–74. Bibcode: 2012Natur.489...57T. doi:10.1038/nature11247.PMC 3439153.PMID 22955616.
- Costa, Fabricio (2012). "7 Non-coding RNAs, Epigenomics, and Complexity in Human Cells". In Morris, Kevin V. *Non-coding RNAs and Epigenetic Regulation of Gene Expression: Drivers of Natural Selection*. Caister Academic Press. ISBN 1904455948.
- Carey, Nessa (2015). *Junk DNA: A Journey Through the Dark Matter of the Genome*. Columbia University Press. ISBN 9780231170840.
- Robin McKie (24 February 2013). "Scientists attacked over claim that 'junk DNA' is vital to life". *The Observer*.
- Sean Eddy (2012) The C-value paradox, junk DNA, and ENCODE, *CurrBiol* 22(21):R898–R899.
- Doolittle, W. Ford (2013). "Is junk DNA bunk? A critique of ENCODE". *Proc Natl AcadSci USA*110 (14): 5294–5300. Bibcode:2013PNAS..110.5294D. doi: 10.1073/pnas.1221376110. PMC 3619371. PMID 23479647.
- Palazzo, Alexander F.; Gregory, T. Ryan (2014). "The Case for Junk DNA". *PLoS Genetics*10 (5): e1004351. doi: 10.1371/journal.pgen.1004351.ISSN 1553- 7404.
- Dan Graur, Yichen Zheng, Nicholas Price, Ricardo B. R. Azevedo1, Rebecca A. Zufall and EranElhaik (2013). "On the immortality of television sets: "function" in the human genome according to the



- evolution-free gospel of ENCODE" (PDF). *Genome Biology and Evolution* 5 (3): 578–90. doi:10.1093/gbe/evt028.3622293. PMID 23431001.
11. Ponting, CP; Hardison, RC (2011). "What fraction of the human genome is functional?". *Genome Research* 21: 1769–1776. doi: 10.1101/gr.116814.110. PMC 3205562. PMID 21875934.
  12. Kellis, M.; et al. (2014). "Defining functional DNA elements in the human genome". *PNAS* 111 (17): 6131–6138. Bibcode: 2014PNAS..111.6131K. doi:10.1073/pnas.1318948111. PMC 4035993. PMID 24753594.
  13. Chris M. Rands, Stephen Meader, Chris P. Ponting and Gerton Lunter (2014). "8.2% of the Human Genome Is Constrained: Variation in Rates of Turnover across Functional Element Classes in the Human Lineage". *PLoS Genet* 10 (7): e1004525. doi:10.1371/journal.pgen.1004525. PMC 4109858. PMID 25057982.
  14. Mattick JS, Dinger ME (2013). "The extent of functionality in the human genome". *The HUGO Journal* 7 (1): 2. doi:10.1186/1877-6566-7-2.
  15. Morris, Kevin, ed. (2012). *Non-Coding RNAs and Epigenetic Regulation of Gene Expression: Drivers of Natural Selection*. Norfolk, UK: Caister Academic Press. ISBN 1904455948.
  16. Elgar G, Vavouri T; Vavouri (July 2008). "Tuning in to the signals: noncoding sequence conservation in vertebrate genomes". *Trends Genet.* 24 (7): 344–52. doi:10.1016/j.tig.2008.04.005. PMID 18514361.
  17. Gregory TR, Hebert PD; Hebert (April 1999). "The modulation of DNA content: proximate causes and ultimate consequences". *Genome Res.* 9 (4): 317–24. doi: 10.1101/gr.9.4.317 (inactive 2015-02-01). PMID 10207154.
  18. Wahls, W.P.; et al. (1990). "Hypervariable minisatellite DNA is a hotspot for homologous recombination in human cells". *Cell* 60 (1): 95–103. doi:10.1016/0092-8674(90)90719-UPMID 2295091.
  19. Waterhouse, Peter M.; Hellens, Roger P. (25 March 2015). "Plant biology: Coding in non-coding RNAs". *Nature* 520 (7545): 41–42. doi:10.1038/nature14378.
  20. Li M, Marin-Muller C, Bharadwaj U, Chow KH, Yao Q, Chen C; Marin-Muller; Bharadwaj; Chow; Yao; Chen (April 2009). "MicroRNAs: Control and Loss of Control in Human Physiology and Disease". *World J Surg* 33 (4): 667–84. doi:10.1007/s00268-008-9836-x. PMC 2933043. PMID 19030926.
  21. Visel A, Rubin EM, Pennacchio LA (September 2009). "Genomic Views of Distant-Acting Enhancers". *Nature* 461 (7261): 199–205. Bibcode:2009Natur.461..199V. doi: 10.1038/nature08451. PMC 2923221. PMID 19741700.
  22. Nielsen H, Johansen SD; Johansen (2009). "Group I introns: Moving in new directions". *RNA Biol* 6 (4): 375–83. doi:10.4161/rna.6.4.9334. PMID 19667762.
  23. Zheng D, Frankish A, Baertsch R; et al. (June 2007). "Pseudogenes in the ENCODE regions: Consensus annotation, analysis of transcription, and evolution". *Genome Res.* 17 (6): 839–51. doi:10.1101/gr.5586307. PMC 1891343. PMID 17568002.
  24. Marshall CR, Raff EC, Raff RA; Raff; Raff (December 1994). "Dollo's law and the death and resurrection of genes". *Proc. Natl. Acad. Sci. U.S.A.* 91 (25): 12283–7. Bibcode:1994PNAS...9112283M. doi: 10.1073/pnas.91.25.12283. PMC 45421. PMID 7991619.
  25. Tutar, Y. (2012). "Pseudogenes". *Comp Funct Genomics* 2012: 424526. doi:10.1155/2012/424526. PMC 3352212. PMID 22611337.
  26. Petrov DA, Hartl DL; Hartl (2000). "Pseudogene evolution and natural selection for a compact genome". *J. Hered.* 91 (3): 221–7. doi:10.1093/jhered/91.3.221. PMID 10833048.
  27. Ponicsan SL, Kugel JF, Goodrich JA; Kugel; Goodrich (February 2010). "Genomic gems: SINE RNAs regulate mRNA production". *Current Opinion in Genetics & Development* 20 (2): 149–55. doi:10.1016/j.gde.2010.01.004. PMC 2859989. PMID 20176473.
  28. Häsler J, Samuelsson T, Strub K; Samuelsson; Strub (July 2007). "Useful 'junk': Alu RNAs in the human transcriptome". *Cell. Mol. Life Sci.* 64 (14): 1793–800. doi:10.1007/s00018-007-7084-0. PMID 17514354.
  29. Walters RD, Kugel JF, Goodrich JA; Kugel; Goodrich (Aug 2009). "InvAluable junk: the cellular impact and function of Alu and B2 RNAs". *IUBMB Life* 61 (8): 831–7. doi:10.1002/iub.227. PMC 4049031. PMID 19621349.
  30. Nelson, PN.; Hooley, P.; Roden, D.; DavariEjtehad, H.; Rylance, P.; Warren, P.; Martin, J.; Murray, PG. (Oct 2004). "Human endogenous retroviruses: transposable elements with potential?". *Clin Exp Immunol* 138 (1): 1–9. doi:10.1111/j.1365-2249.2004.02592.x. PMC 1809191. PMID 15373898.
  31. International Human Genome Sequencing Consortium (February 2001). "Initial sequencing and analysis of the human genome". *Nature* 409 (6822): 879–888. Bibcode: 2001Natur.409..860L. doi:10.1038/35057062. PMID 11237011.
  32. Piegu, B.; Guyot, R.; Picault, N.; Roulin, A.; Sanyal, A.; Sanyal, A.; Kim, H.; Collura, K.; et al. (Oct 2006). "Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice". *Genome Res* 16 (10): 1262–9. doi:10.1101/gr.5290206.1581435. PMID 16963705.
  33. Hawkins, JS.; Kim, H.; Nason, JD.; Wing, RA.; Wendel, JF. (Oct 2006). "Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*". *Genome Res* 16 (10): 1252–61. doi:10.1101/gr.5282906.1581434. PMID 16954538.

34. Ehret CF, De Haller G; De Haller (1963). "Origin, development, and maturation of organelles and organelle systems of the cell surface in Paramecium". *Journal of Ultrastructure Research*. 9 Supplement 1: 1, 3–42. doi:10.1016/S0022-5320(63)80088-X. PMID 14073743.
35. Dan Graur, *The Origin of Junk DNA: A Historical Whodunnit*
36. Gregory, T. Ryan, ed. (2005). *The Evolution of the Genome*. Elsevier. pp. 29–31. ISBN 0123014638. Comings (1972), on the other hand, gave what must be considered the first explicit discussion of the nature of "junk DNA," and was the first to apply the term to all noncoding DNA.; "For this reason, it is unlikely that any one function for noncoding DNA can account for either its sheer mass or its unequal distribution among taxa. However, dismissing it as no more than "junk" in the pejorative sense of "useless" or "wasteful" does little to advance the understanding of genome evolution. For this reason, the far less loaded term "noncoding DNA" is used throughout this chapter and is recommended in preference to "junk DNA" for future treatments of the subject."
37. Ohno, Susumu (1972). H. H. Smith, ed. *So Much "junk" DNA in Our Genome*. Gordon and Breach, New York. pp. 366–370. Retrieved 2013-05-15.
38. Doolittle WF, Sapienza C; Sapienza (1980). "Selfish genes, the phenotype paradigm and genome evolution". *Nature* 284 (5757): 601–603. Bibcode:1980Natur.284..601D. doi:10.1038/284601a0. PMID 6245369.
39. Another source is genome duplication followed by a loss of function due to redundancy.
40. Orgel LE, Crick FH; Crick (April 1980). "Selfish DNA: the ultimate parasite". *Nature* 284 (5757): 604–7. Bibcode:1980Natur.284..604O. doi:10.1038/284604a0. PMID 7366731.
41. Khajavinia A, Makalowski W; Makalowski (May 2007). "What is "junk" DNA, and what is it worth?". *Scientific American* 296 (5): 104. doi: 10. 1038/scientificamerican0307-104. PMID 17503549. The term "junk DNA" repelled mainstream researchers from studying non coding genetic material for many years
42. Biémont, Christian; Vieira, C (2006). "Genetics: Junk DNA as an evolutionary force". *Nature* 443 (7111): 521–4. Bibcode:2006Natur.443..521B. doi:10.1038/443521a. PMID 17024082.
43. Palazzo, Alexander F.; Lee, Eliza S. (2015). "Non-coding RNA: what is functional and what is junk?". *Frontiers in Genetics* 6: 2. doi: 10.3389/fgene.2015.00002. ISSN 1664-8021. PMID 25674102.
44. Ludwig MZ (December 2002). "Functional evolution of noncoding DNA". *Current Opinion in Genetics & Development* 12 (6): 634–9. doi:10.1016/S0959-437X(02)00355-6. PMID 12433575.
45. Cobb J, Büsst C, Petrou S, Harrap S, Ellis J; Büsst; Petrou; Harrap; Ellis (April 2008). "Searching for functional genetic variants in non-coding DNA". *Clin. Exp. Pharmacol. Physiol.* 35 (4): 372–5. doi:10.1111/j.1440-1681.2008.04880.x. PMID 18307723.
46. E Khurana; et al. (April 2013). "Integrative annotation of variants from 1092 humans: application to cancer genomics". *Science* 342 (6154): 372–5. doi: 10.1126/science.1235587. PMC 3947637. PMID 24092746.
47. Lu, Yi-Fan; Mauger, David M.; Goldstein, David B.; Urban, Thomas J.; Weeks, Kevin M.; Bradrick, Shelton S. (4 November 2015). "IFNL3 mRNA structure is remodeled by a functional non-coding polymorphism associated with hepatitis C virus clearance". *Scientific Reports* 5: 16037. doi:10.1038/srep16037. PMID 26531896.
48. Grünewald, Thomas G P; Bernard, Virginie; Gilardi-Hebenstreit, Pascale; Raynal, Virginie; Surdez, Didier; Aynaud, Marie-Ming; Mirabeau, Olivier; Cidre-Aranaz, Florencia; Tirode, Franck. "Chimeric EWSR1-FLI1 regulates the Ewing sarcoma susceptibility gene EGR2 via a GGAA microsatellite". *Nature Genetics* 47 (9): 1073–1078. doi:10.1038/ng.3363. PMC 4591073. PMID 26214589.
49. Subirana JA, Messeguer X; Messeguer (March 2010). "The most frequent short sequences in non-coding DNA". *Nucleic Acids Res.* 38 (4): 1172–81. doi:10.1093/nar/gkp1094. PMC 2831315. PMID 19966278.
50. S. E. Ahnert; T. M. A. Fink (2008). "How much non-coding DNA do eukaryotes require?" (PDF). *J. Theor. Biol.* 252 (4): 587–592. doi:10.1016/j.jtbi.2008.02.005. PMID 18384817.
51. Smith MA; et al. (June 2013). "Widespread purifying selection on RNA structure in mammals". *Nucleic Acids Research* 41 (17): 8220–8236. doi:10.1093/nar/gkt596. PMC 3783177. PMID 23847102.
52. Dileep, V (2009). "The place and function of non-coding DNA in the evolution of variability". *Hypothesis* 7 (1): e7. doi:10.5779/hypothesis.v7i1.146.
53. Carroll, Sean B.; et al. (May 2008). "Regulating Evolution". *Scientific American* 298 (5): 60–67. doi: 10.1038/scientificamerican0508-60. PMID 18444326.
54. Stojic, L. "Transcriptional silencing of long noncoding RNA GNG12-AS1 uncouples its transcriptional and product-related functions". *nature.com. Nature*. Retrieved 21 Feb 2016.
55. Callaway, Ewen (March 2010). "Junk DNA gets credit for making us who we are". *New Scientist*.
56. "Plagiarized Errors and Molecular Genetics", *talkorigins*, by Edward E. Max, M.D., Ph.D.
57. Balakirev ES, Ayala FJ; Ayala (2003). "Pseudogenes: are they "junk" or functional DNA?". *Annu. Rev. Genet.* 37: 123–51. doi:10.1146/annurev.genet.37.040103.103949. PMID 14616058.



58. C.-K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, H. E. Stanley; Buldyrev, SV; Goldberger, AL; Havlin, S; Sciortino, F; Simons, M; Stanley, HE (1992). "Long-range correlations in nucleotide sequences". *Nature* 356 (6365): 168–70. Bibcode:1992Natur.356..168P. doi:10.1038/356168a0. PMID 1301010.
59. W. Li and, K. Kaneko; Kaneko, K (1992). "Long-Range Correlation and Partial  $1/f^{\alpha}$  Spectrum in a Non-Coding DNA Sequence" (PDF). *Europhys. Lett* 17 (7): 655–660. Bibcode:1992EL.....17..655L. doi:10.1209/02955075/17/7/014.
60. S. V. Buldyrev, A. L. Goldberger, S. Havlin, R. N. Mantegna, M. Matsu, C.-K. Peng, M. Simons, and H. E. Stanley; Goldberger, A.; Havlin, S.; Mantegna, R.; Matsu, M.; Peng, C.-K.; Simons, M.; Stanley, H. (1995). "Long-range correlations properties of coding and noncoding DNA sequences: GenBank analysis". *Phys. Rev. E* 51 (5): 5084–5091. Bibcode: 1995PhRvE..51.5084B. doi:10.1103/PhysRevE.51.5084.