MHR

J. C. BIRO AND N. BAROUKH [2005] MED HYPOTHESES RES 2: 303-313.

PREDICTION OF GENOME-WIDE ASSOCIATIVE REGULATORY ROLE OF SHORT AND LONG INTERSPERSED NUCLEOTIDE ELEMENTS (SINE AND LINE)

J. C. BIRO* AND N. BAROUKH

KAROLINSKA INSTITUTE, STOCKHOLM, SWEDEN (J.C.B.) AND GENOME SCIENCES DEPARTMENT, LAWRENCE BERKELEY NATIONAL LABORATORY, BERKELEY, CA, USA (N.B.)

REVIEW

ABSTRACT. REPETITIVE ELEMENTS represent the same or similar biological information many thousands times dispersed in the genome. This property makes them simultaneously suitable for associative regulation of the entire genome. Alu elements (SINE) often exonize, becoming transcribed as part of a larger gene, and the resulting mRNA is complementary to the genomic DNA at ~1.4 million sites. LINE retro-transposons are able to move and leave a signature at thousands of different genes. Many of them have multiple transcriptions factor (TF) binding sites and probably function as promoters, which indicate that some of the single TFs might activate large clusters of previously dedicated genes. Simple repeats have low information content, but still enough to influence quantitative parameters of biological regulation.

Special attention has been paid to a LINE element, called NAD (next to the APOA conserved domain) located in the human APOA1/C3/A4/A5 gene cluster and highly conserved, even in the mouse genome. NAD shows all necessary bioinformatical signatures of a promoter and is represented at 260 places in the human genome, but non-functional in transgenic mice. We interpret these results in favor of existence of species-specific, conditional promoters (in contrast to the conventional cis-regulatory elements).

*ADDRESS ALL CORRESPONDENCE TO: DR. J.C. BIRO, 88 HOWARD, NO. 1205, SAN FRANCISCO, 94195 CA, USA. E-MAIL: jan.biro@sbcglobal.net

• MEDICAL HYPOTHESES AND RESEARCH • THE JOURNAL FOR INNOVATIVE IDEAS IN BIOMEDICAL RESEARCH •

1. INTRODUCTION

Approximately 25% of the human genome consists of repetitive sequences. They vary in size, distribution, number and complexity [1] (TABLE 1). Some are repetitions of only a few residues and have almost no sequence complexity at all (like the telomer sequences). Others are obviously carriers of complex genetic information and often behave like parasites in the host genome (like retro-transposons). Conventionally they are treated as "junk" that accumulated during evolution in higher organisms. Many bioinformatics tools are installed by default to filter and remove repeats and low complexity sequences before performing any analyses on the rest of sequences.

However, the success of Human Genome Project, as well as other whole genome sequencing efforts are providing an increasing volume of evidence for a significant biological role of the repetitive sequences. This article is aimed to provide some arguments for the importance of these sequences and a theory on their novel functions.

2. MATERIALS AND METHODS

SINE (short interspersed nuclear element) and LINE (long interspersed nuclear element) sequences were taken from the EMBL and RepBase databases [2,3]. They were translated into the six possible translation frames using codon translation TABLE 1 [4].

Statistical analyses of the amino acid composition of the translated (protein) sequences was performed by SAPS [5,6]. ClustalW method was used for Multiple Sequence Alignments, MSA [7,8] and psi-blast method for sequence similarity searching [9,10]. Comparative genomic data were mined by the VISTA and ENSEMBL genome browsers [11-14]. Transcription factor (TF) data were collected by rVISTA (regulatory) using default matrices and parameters [15,16].

The creation and properties of transgenic mice, containing human Apo-lipoprotein A cluster (APOA1/C3/A4/A5 gene cluster) is described elsewhere [17-19]. Students' t-test was used for statistical evaluation of the data obtained.

3. RESULTS AND DISCUSSION

Most of our knowledge about nucleic acid structure was accumulated since 1953 when Watson and Crick won the scientific race against Rosalin Franklin [20], Mauritz Wilkins [21,22], Robert Corey and Linus Pauling [23], with the publication of a dominant model for the double helical structure of DNA [24,25]. For a half century only fragments of the DNA from different organisms were available for studies, because whole genome sequencing was not possible. However this was dramatically changed when Craig Venter introduced the shotgun sequencing technique and the concurrence between the Human Genome Sequencing Project and Celera company accelerated genome discovery and supplied complete genome primary data (sequence) for many organisms including human [26,27]. Now, when it begins to be possible to see the "big picture" we are forced to incite that many of our primary concepts of the gene structure and function are wrong or primitive. For example, mapping of the human genome clearly showed that both DNA strands are transcribed [28] and many mRNA never became translated, having instead a regulatory role [29]. Similarly, revolutionary revision of the biological role of repetitive sequences is also in progress.

3.1. SINE

Alu repetitive elements are found in ~1.4 million

TABLE 1. DISTRIBUTION OF REPEATED DNA SEQUENCES.

Species	Single copy (%)	Sequence distribution 10-1000 copies (%)	$10^5 – 10^6$ copies (%)
Bacteria	99.7		
Mouse	60	25	10
Human	70	13	8
Cotton	61	27	8
Corn	30	40	20
Wheat	10	83	4
Arabidopsis	55	27	10

Source: McClean [1997].

copies in the human genome, comprising more than one-tenth of it. Numerous studies describe exonizations of Alu elements, that is, splicing-mediated insertions of parts of Alu sequences into mature mRNAs [30]. There are many examples for this, and a database, called AluGene, is established to list Alu elements incorporated within protein-coding genes [31,32]. There are 8 main Alu classes, each being about 300 bases long, and the average distance between different classes is low (<6%). We translated the Alu sequences into the 6 possible translation frames. Each frames contained many stop signals, which might explain while only fragments of Alu are able to exonize and not entire Alu sequences. Statistical analyses of the physico-chemical properties of translated sequences showed, that all frames in every Alu class-specific consensus sequences have positive average charge (FIG. 1). Similarity search of non-redundant protein databases (NCBI) with translated Alu-queries and using the psi-blast method showed significant and extensive homology to nuclear regulatory proteins. (TABLE 2).

On the bases of the literature data as well as our recent and previous research, we wish to propose hypotheses to explain the biological role of the Alu repeats. The large number of short repeats in the genome indicates to us that they are fundamental but non-specific elements of the genome. They are double stranded and mostly silent components of the dsDNA. However a few active regulatory genes express Alu-like domains. The Alu-like mRNA and the Alu-like proteins interact with the DNA (the Alu-like mRNA is complementary to one Alu-DNA strand). We suggest that: (i) A few slightly different Alu-like mRNAs can cower an entire Alu sequence;

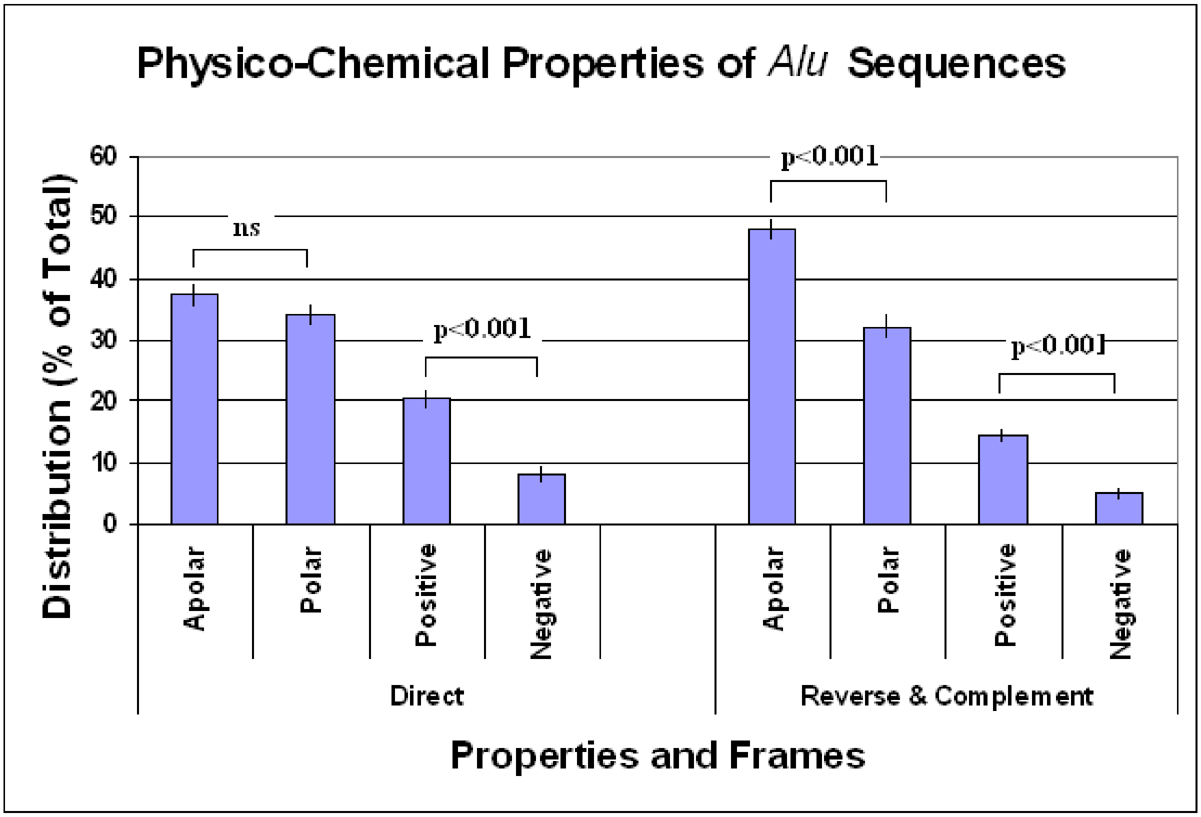

FIGURE 1. PHYSICO-CHEMICAL PROPERTIES OF ALU SEQUENCES. Alu sequences (one of each classes) were translated (each into the 6 possible frames) and the physico-chemical properties of the residues were statistically analyzed. Each bar represents MEAN ± S.E.M. (N = 24).

TABLE 2. *Alu*-HOMOLOGS FOUND BY PSI-BLAST.

Subject		Query			Homology parameters				
Name	Accession Nr	Frame	Start	End	Length	Identity %	Similarity %	Gaps %	E-value
B-cell growth factor	AAB02649.1	D1	6	43	38	71	78	0	2.00E-08
RNA polymerase I	AAC39892.1	D1	6	29	24	66	79	0	9.00E-02
CREBL1 protein	AAH08394.1	D2	2	36	35	77	82	0	2.00E-06
Leucine zipper-like protein	AAD47042.1	D2	6	36	31	77	80	0	3.00E-03
NF-kappa-B2	A57034	D3	3	46	44	54	65	0	1.00E-02
Zinc finger protein	AAB04930.1	D3	18	49	32	53	65	0	3.00E-01
Topoisomerase II α-4	AAG13405.1	RC1	57	94	38	78	84	0	3.00E-10
Krueppel zinc finger protein	AAB86596.1	RC1	62	99	38	78	86	0	6.00E-10
v-myb oncogene	AACAE55174.1	RC1	57	99	43	79	81	0	2.00E-09
c-myb13A	CAF04484.1	RC2	16	65	50	72	72	2	3.00E-12
Topoisomerase II α-4	AAG13405.1	RC2	5	63	39	49	59	7	2.00E-09

(ii) the number of Alu-like sequences present in the nuclei might reach a concentration when they destabilize (dissociate) the dsDNA; (iii) when it happens, it happens at 1.4 million DNA sites at the same time; and (iv) that will trigger mitosis and the cell will divide (FIG. 2).

3.2. LINE

Long interspersed nuclear elements (LINEs) are endogenous mobile genetic elements that have dispersed and accumulated in the genomes of higher eukaryotes via germline transposition, with up to 100,000 copies in mammalian genomes. In humans, LINEs are the major source of insertional mutagenesis, being involved in both germinal and somatic mutant phenotypes.

The LINE elements are much longer than Alu but there are much fewer identical copies present in the genome. The HAL1 class sequences are, for instance, about 1,700 residues long and there are about 1,400 almost identical copies present in the human genome. The LINE subclasses are very different from each other.

A remarkable feature of LINE retro-transposons is that they transpose through the reverse transcription of their own sequence [33]. They are a significant source of genome diversity [34]. The seemingly random migration and dispersion of LINE elements might reach invasive proportions and in the 80s some scientists regarded them as genomic parasites. This view led to the concept of so-called "selfish DNA" and the “fluidity of the genome”. However, accumulating evidence suggests that the distribution of LINE is not random [35]. They are all believed to be generally suppressed, but there are examples of selective regulation by surrounding and distantly located genes [36].

It is now widely accepted, that retro-transposons do play a significant role in the evolution of mammalian gene expression [37] and many new investigations indicate that they even actively regulate the actual physiological expression of some genes [38]. They not only have their own promoters but in addition are promoters of other genes [39-41].

We have a long history of a desire to understand the rules of gene regulation generally and the apo-lipoprotein gene cluster particularly [19,42,43]. Our laboratory in Berkeley is recently characterizing the structure and function of the human APOA1/C3/A4/A5 gene cluster [44] and created a transgenic mice containing this cluster and expressing it’s genes [18,44]. The cluster contains an 804 residue long conserved noncoding sequence (75% human/mice identity). This sequence, called NAD (next to APO-conserved domain) is located

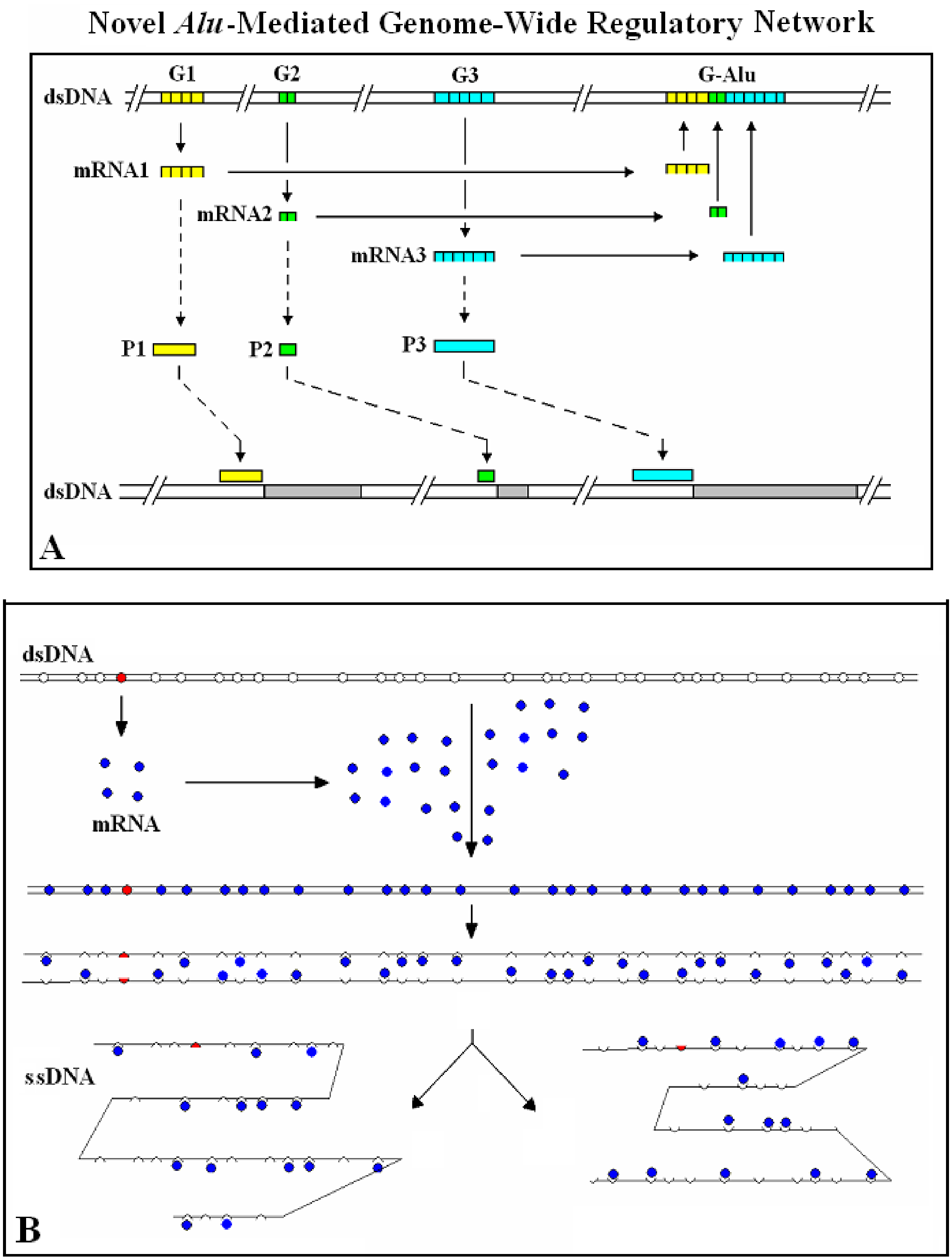

FIGURE 2. NOVEL ALU-MEDIATED GENOME-WIDE REGULATORY NETWORK. A. Three different genes in the genome (G1, G2 and G3) express parts of a complete Alu sequence. The exonized Alu sequences (mRNA1-2-3) are complementary to the Alu gene (G-Alu) and bind to it. They are even translated into nuclear regulatory proteins (P1-2-3) and contribute to the regulation of other, non-Alu genes in an other double stranded (ds) DNA. B. The intranuclear concentration of exonized Alu-containing mRNAs (dots) are complementary to the numerous Alu repeats in the genome (dsDNA), specifically binds to them and dissociate the entire dsDNA into two single stranded (ss) DNA strands.

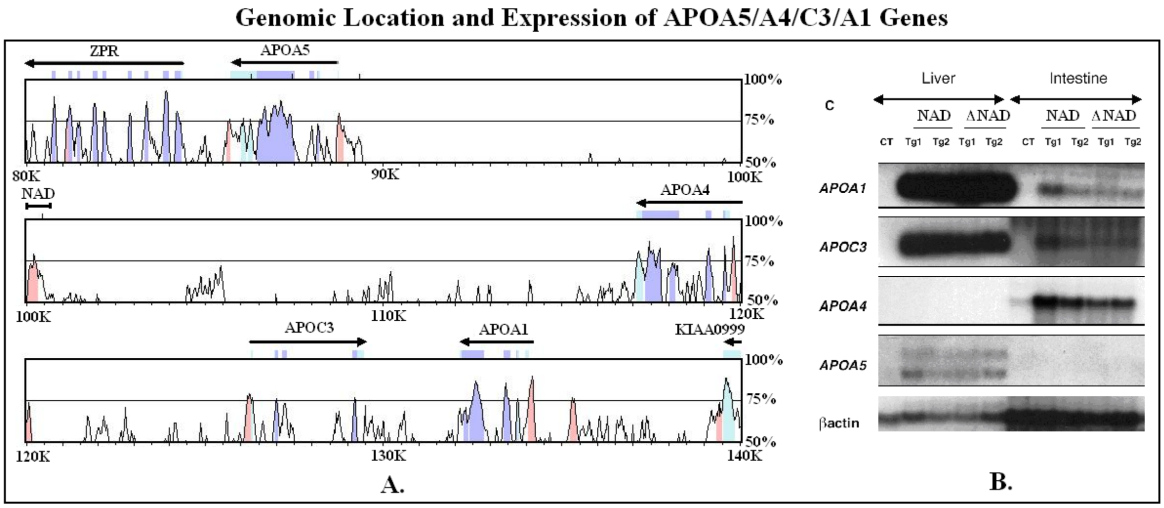

FIGURE 3. GENOMIC LOCATION AND EXPRESSION OF APOA5/A4/C3/A1 GENES.

A. A human-mouse VISTA plot displaying the level of genomic sequence conservation. Arrows indicate known genes and their orientation with each exon depicted by a box (gene names are indicated above each arrow). Human sequence is represented horizontally (x-axis) and the percent similarity with the mouse sequence is plotted vertically (y-axis). NAD: Next to APOA conserved domain.

B. RNA analyses of transgenic mice (Tg) containing the conserved noncoding sequence (NAD) compared to transgenic mice lacking the element (ΔNAD) and wild-type control mouse (CT). Liver and intestine total RNA were prepared and hybridized with human-specific probes for APOA1, APOC3, APOA4 and APOA5. Mouse β-actin was used as an internal control.

~13 kb upstream to APOA5 gene, It was a given candidate to be an APOA5 promoter. But it is a HAL1 type LINE, the only conserved LINE in the cluster. Enthusiastically, but not without hesitation, we deleted this sequence just to learn that this is not a cis-promoter: excising of the sequence had no effect on the expression of any APOA genes in any tissues (FIG. 3).

However, in a larger perspective, the NAD sequence still remains interesting. It contains numerous novel transcriptions factor binding site, similarly to other LINEs (FIG. 4). It is present in all, but Y chromosomes, in 260 slightly different copies. This slight sequence variation causes significant differences in the pattern of the potential transcription factor binding sites (TABLE 3). It is a difficult task to correctly predict TF binding sites because these sequences are short, not giving enough specificity. Therefore the number of binding sites is regularly overestimated. Our statistical analyses indicate that only about 30% of the predictions are correct but that is more than enough to claim that NAD and other LINE sequences might be promoter-like regulatory elements.

An additional interesting feature of the NAD and HAL1 sequences is that they preferentially locate close to a TF-gene. About a third of NAD and HAL1 - like sequences in the human genome are located next to or at least on the same contig as a TF gene. ALUJ repeats (used as controls) do not show the same preferential location (FIGURE 5).

Based on recent literature and also our own experience with the NAD, we propose that the LINEs are novel genome-wide, associative, parallel regulators and coordinators of the functions of major gene clusters, and it may play an essential role in determining the functional and morphological phenotype of a species.

In this theory, the phylogenesis of genes and functions goes in two parallel ways: (i) The development of a species specific set of genes, and (ii) the development of the species specific way to coordinate the genes (i.e., their functions).

The first big surprise caused by the HUGO and other whole genome sequencing projects was the

TABLE 3. NAMES AND COPY NUMBERS OF PREDICTED TF BINDING SITES IN THE MOST NAD-LIKE SEQUENCES IN EACH HUMAN CHROMOSOME.

No.	Chromosome	20	15	21	22	12	4	1	8	9	16	X	11	19	3	18	10	2	7	6	5	17	14	13	NAD	Summa
	TF name																									row
1	CP2											1														1
2	E2F																					1				1
3	E2F_Q3																					1				1
4	E2F1DP1																					1				1
5																						1				1
…																										…
76	HNF1_Q6	1					2	2	2			3	3	3	4	6	3	4	5	1	3	2	7	9	7	67
77	MEF2_Q6			1		2	1	3		1		1		6	5		2	8	4	12	65		8		12	72
78	CDC5		6	1	1	1	8			4	2	2	8	1	2	3	4	5	3	8	3	2	4		5	73
79	OCT1_Q6			1	1		3		2	4	3		6		2	6	1	2	4	3	5	10	3	15	2	73
80	GATA4_Q3			1	2	2		2		4	6	3	2	4	3	8	4	2			7	4	1	15	6	76
	Summa column	11	13	14	14	17	20	21	21	25	27	29	32	32	38	43	45	46	47	49	53	57	67	93	94	908

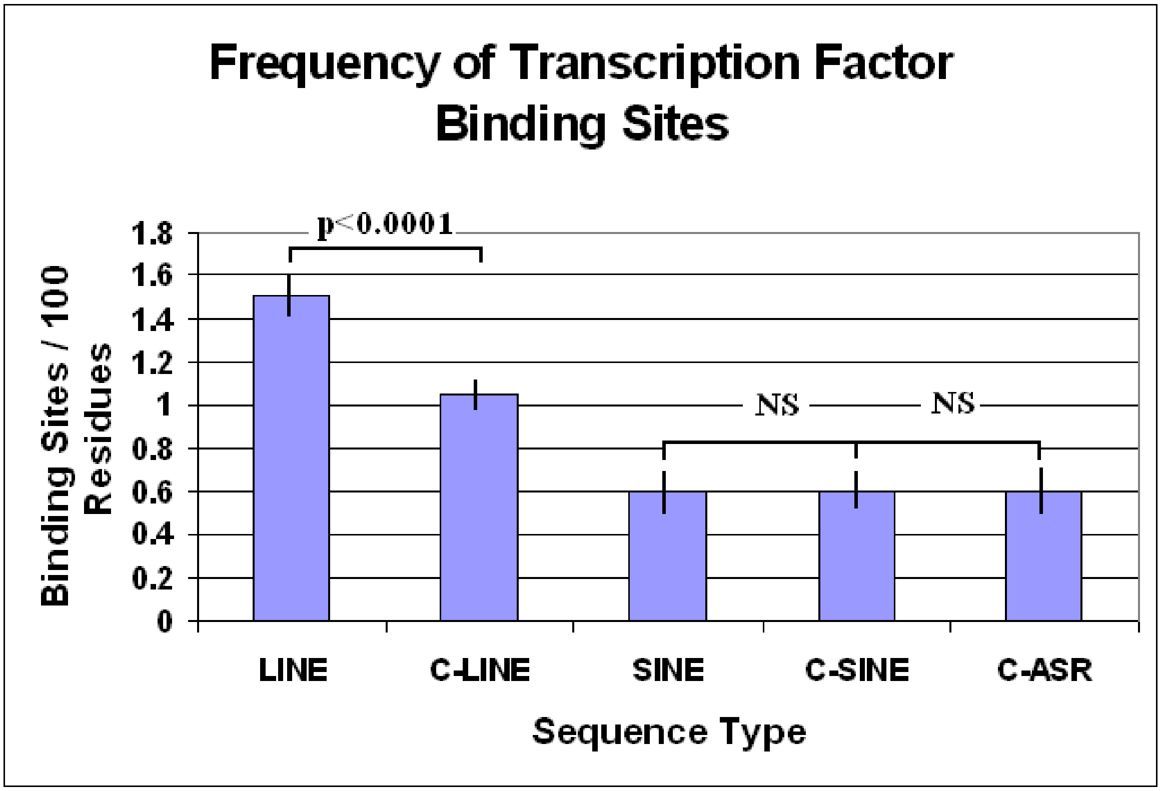

FIGURE 4. FREQUENCY OF TRANSCRIPTION FACTOR (TF) BINDING SITES. Potential TF binding sites in LINE and SINE sequences were predicted by rVISTA. Shuffled sequences and artificial random sequences (ASR) served as controls (C). Each bar represents MEAN ± S.E.M. (N = 5-8).

discovery that the number of human genes is unexpectedly small and that there is no reasonable correlation between the number of genes and the complexity of an organism. Fortunately the gene-splicing variation increases during phylogenesis and it might explain the larger complexity of a younger species. We are not really surprised over the relatively small number and small inter-species variation of the genes, for we believe that the genes are tools, and living in the same biological

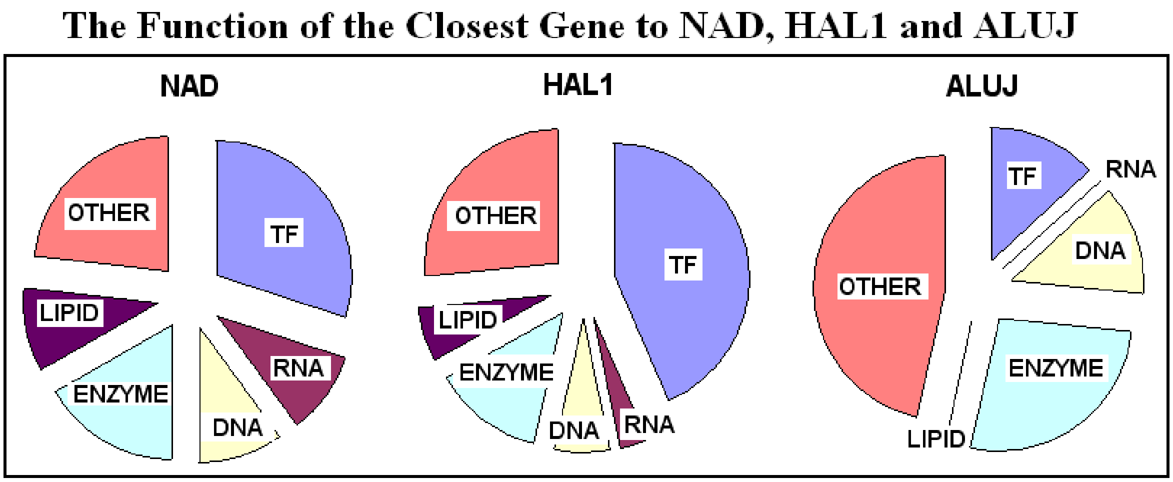

FIGURE 5. THE FUNCTION OF THE CLOSEST GENES TO NAD, HAL1 AND ALUJ SEQUENCE. Thirty most similar sequences to NAD, HAL1 and AluJ were located in the human genome with ENSEML genome browser and the function of the closest genes was recorded. The circles indicate the proportion of transcription factors (TF), genes related to DNA, RNA, lipid metabolism, enzymes and genes with other functions.

environment requires very similar biochemical tools. We think that the real difference between the species arises from the different efficiency and sophistication to use these tools. The source of the major inter-species differences (phenotypes) seems to be located mostly in the introns and much less in the exons. (Have you ever been in a French, German, English, or Italian kitchen? They all process very similar ingredients with the same kinds of equipment, but the cook books are different and that makes the difference in the taste of the food). The "cook book" of the individual chemistry is probably located in the introns.

In our theory of genome-wide associative regulation of genes the large diverse family of mobile LINEs has a central role (FIG. 6). During the evolution, one or more LINEs successively "jump" to a subset of genes which are simultaneously active during a longer period of time and stable conditions, physically associate to those genes and add a common signature to every member of that particular subset of genes. In that way gene clusters develop, each having their own specific LINE signature, which indicates that under a given particular condition they perform together. One gene (a biological tool) can belong to many different clusters and may have many different LINE signatures. Activating a LINE by one or a combination of, say, transcriptions factors will activate the entire gene cluster and start to perform a multi-gene action. In that way, a gene cluster or a combination of clusters will give rise to a species specific morphological or biochemical phenotype.

3.3. SIMPLE REPEATS

Very short oligonucleotides (3-6 residues) may repeat several thousand times. The information content of these simple repeats is low, therefore they cannot specify any distinct gene quality. They are, however, related to some quantitative aspects of gene function.

New types of neurological disorders were found in the 1990s. These types of the diseases are caused by an expansion of repetitive three bases in the causative gene. This is called a triplet repeat disease. For example, fragile X syndrome is caused by an expansion of CGG repeat, Huntington's disease by CAG repeat, Friedreich's ataxia by GAA repeat and myotonic dystrophy by CTG repeat. The symptoms of these diseases become more severe, and onset is earlier with each successive generation. Of these diseases, the one caused by an extension of CAG repeat is specially called CAG repeat or polyglutamine [poly(Q)] disease. The number of CAG repeat is about 20 in normal individuals, but is

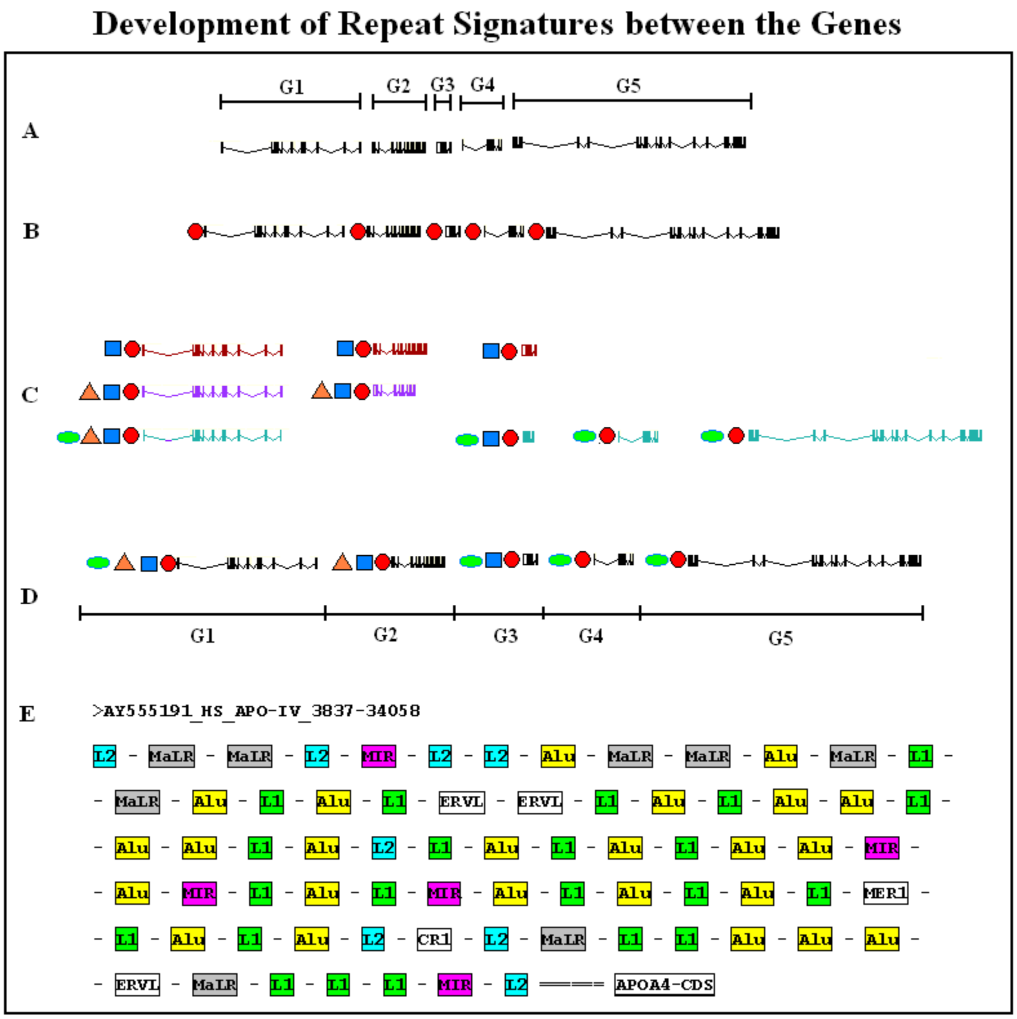
Development of Repeat Signatures between the Genes
A
G1
G2
G3
G4
G5
B
C
D
G1
G2
G3
G4
G5
E
>AY555191_HS_APO-IV_3837-34058
L2 - MaLR - MaLR - L2 - MIR - L2 - L2 - Alu - MaLR - MaLR - Alu - MaLR - L1 -
- MaLR - Alu - L1 - Alu - L1 - ERVL - ERVL - L1 - Alu - L1 - Alu - Alu - L1 -
- Alu - Alu - L1 - Alu - L2 - L1 - Alu - L1 - Alu - L1 - Alu - Alu - MIR -
- Alu - MIR - L1 - Alu - L1 - MIR - Alu - L1 - Alu - L1 - Alu - L1 - MER1 -
- L1 - Alu - L1 - Alu - L2 - CR1 - L2 - MaLR - L1 - L1 - Alu - Alu - Alu -
- ERVL - MaLR - L1 - L1 - L1 - MIR - L2 ==== APOA4-CDS

FIGURE 6. DEVELOPMENT OF REPEAT SIGNATURE BETWEEN THE GENES. Five genes (G1-5) are physically linked to each other at the beginning of their development (A) with almost no introns between them. They are expressed together and receive a common signature (circle) at early developmental stadium (B). However during later differentiation and evolution they might be expressed in subsets and receive a specific subset (cluster) specific signature (squares, triangles, ovals) (C). Finally a complex pattern of context-signatures develop and expands the inter-genomic space (D). A simple gene became highly decorated by functional and developmental signatures as it is illustrated on a real human gene (E). The APOA4-coding sequence (CDS) is proceeded by 24 SINE (Alu) and 48 LINE signatures.

40 or more in patients with the disease. Huntington's disease (HD) is the most common of these disorders and there is a significant negative linear correlation between the number of CAG repeats in a key protein and the age of onset of this disease [45].

Telomeric repeats are also famous examples of simple repeats. Each time mitosis occurs, the telomeres of the dividing cells get just a bit shorter. Once a cell's telomeres have reached a critically short length, that cell can no longer divide. Its structure and function begin to fail. Some cells even die. The telomeres of humans consist of as many as 2000 repeats of the sequence 5'-TTAGGG-3'. It is estimated that human telomeres lose about 100 base pairs from their telomeric DNA at each mitosis. This represents about 16 TTAGGG repeats. At this rate, after 125 mitotic divisions, the telomeres would be completely gone. Human fibroblast in vitro usually stops dividing already after 50 divisions. There is a positive correlation between the telomere length and the expected number of cell divisions [46]. However this is not a simple issue.

Bacillus Anthracis produce an exosporium glycoprotein which is a structural constituent of the hair-like filaments on the outer layer of the exosporium. This is a collagen-like protein (BclA) which contains an internal collagen like region (CLR) of 19-91 GXX (mostly GPT) repeats and 1–8 copies of (GTP)5GDTGTT. The length of the BclA is responsible for the variation in filament length [47].

The biological role of simple repeats seems to be counting and time-keeping.

Our conclusion is that the gene-centric view of the genome, which was developed in the 50's and was mainly based on data from phages and bacteria, is clearly insufficient to describe the organization and function of higher organisms. The availability of whole genome sequence data makes it possible and necessary to further develop our concepts about the possibilities of biological regulation. It is time to listen to the (repeated) message of the repeats and take seriously that they are "sequences alike" and that that has a meaning.

ACKNOWLEDGEMENTS

The theoretical and most of bioinformatical part of this work was done by J.C.B. while the experimental part entirely by N.B. This research was supported by grants from the Homulus Foundation (J.C.B.) and the American Heart Association (N.B.).

REFERENCES

[1] MCCLEAN P [1997] Eukaryotic chromosome structure. http://www.ndsu.nodak.edu/instruct/mcclean/plsc431/eukarychrom/eukaryo3.htm

[2] JURKA J [2000] Repbase update: A database and an electronic journal of repetitive elements. Trends Genet. 16: 418-420.

[3] REPBASE. http://www.girinst.org/Repbase_Update.html

[4] CODON TRANSLATION TABLES.

http://www.kazusa.or.jp/codon/

[5] BRENDEL V, BUCHER P, NOURBAKHSH I, BLAISDELL BE, KARLIN S [1992] Methods and algorithms for statistical analysis of protein sequences. Proc Natl Acad Sci USA 89: 2002-2006.

[6] SAPS.

http://www.ch.embnet.org/software/SAPS_form.html

[7] HIGGINS D, THOMPSON J, GIBSON T, THOMPSON JD, HIGGINS DG AND GIBSON TJ [1994] CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673-4680.

[8] CLUSTALW. http://www.ebi.ac.uk/clustalw/#

[9] ALTSCHUL SF, MADDEN TL, SCHAFFER AA, ZHANG J, ZHANG Z, MILLER W, LIPMAN DJ [1997] Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucl Acids Res 25: 3389-3402.

[10] PSI BLAST: http://www.ncbi.nlm.nih.gov/BLAST/

[11] MAYOR C, BRUDNO M, SCHWARTZ JR, POLIAKOV A, RUBIN EM, FRAZER KA, PACHTER LS AND DUBCHAK I [2000] VISTA: Visualizing global DNA sequence alignments of arbitrary length. Bioinformatics 16: 1046.

[12] GENOME VISTA. http://gsd.lbl.gov/vista/index.shtml

[13] HUBBARD T, ANDREWS D, CACCAMO M, CAMERON G, CHEN Y, CLAMP M, CLARKE L, COATES G, COX T, CUNNINGHAM F, CURWEN V, CUTTS T, DOWN T, DURBIN R, FERNANDEZ-SUAREZ XM, GILBERT J, HAMMOND M, HERRERO J, HOTZ H, HOWE K, IYER V, JEKOSCH K, KAHARI A, KASPRZYK A, KEEFE D, KEENAN S, KOKOCINSCI F, LONDON D, LONGDEN I, MCVICKER G, MELSOPP C, MEIDL P, POTTER S, PROCTOR G, RAE M, RIOS D, SCHUSTER M, SEARLE S, SEVERIN J, SLATER G, SMEDLEY D, SMITH J, SPOONER W, STABENAU A, STALKER J, STOREY R, TREVANION S, URETA-VIDAL A, VOGEL J, WHITE S, WOODWARK C, BIRNEY E [2004] Ensembl 2005. Nucleic Acids Res 33: D447-D453.

[14] ENSEMBL. http://www.ensembl.org

[15] LOOTS G, OVCHARENKO I, PACHTER L, DUBCHAK I AND RUBIN E [2002] rVISTA for comparative sequence-based discovery of functional transcription factor binding sites. Genome Res 12: 832-839.

[16] rVISTA. http://gsd.lbl.gov/vista/rvista/submit.shtml

[17] BABINET C, MORELLO D AND RENARD JP [1989] Transgenic mice. Genome 31: 938-949.

[18] TALMUD PJ, HAWE1 E, MARTIN S, OLIVIER M, MILLER GJ, RUBIN EM, PENNACCHIO LA AND HUMPHRIES SE [2002] Relative contribution of variation within the APOC3/A4/A5 gene cluster in determining plasma triglycerides. Human Mol Genet 11: 3039-3046.

[19] BAROUKH N, BAUGE E, AKIYAMA J, CHANG J, AFZAL V, FRUCHART JC, RUBIN EM, FRUCHART J AND PENNACCHIO LA [2004] Analysis of apolipoprotein A5, C3, and plasma triglyceride concentrations in genetically engineered mice. Arterioscler Thromb Vasc Biol 24: 1297-1302.

[20] FRANKLIN R AND GOSLING R [1953] Molecular configuration in sodium thymonucleate. Nature 171: 740-741.

[21] FEUGHELMAN M, LANGRIDGE R, SEEDS WE, STOKES AR, WILSON HR, HOOPER CW, WILKINS MH, BARCLAY RK, HAMILTON LD [1955] Molecular structure of deoxyribose nucleic acid and nucleoprotein. Nature 175: 834-838.

[22] SPENCER M, FULLER W, WILKINS MHF AND BROWN GL [1962] Determination of the helical configuration of ribonucleic acid molecules by X-ray diffraction study of crystalline amino-acid-transfer ribonucleic acid. Nature 194: 1014-1020.

[23] PAULING L AND COREY RB [1953] A proposed structure for the nucleic acids. Proc Nat Acad Sci USA 39: 84-97.

[24] WATSON JD AND CRICK FHC [1953] A Structure for Deoxyribose Nucleic Acid. Nature 171: 737-738.

[25] WATSON JD [1980] In: The Double Helix: A Personal Account of the Discovery of the Structure of DNA. Stent G (ed.), New York, Norton.

[26] VENTER JC, ADAMS MD, MYERS EW ET AL. [2001] The sequence of the human genome. Science 292: 291: 1304-1351.

[27] LANDER ES, LINTON LM, BIRREN B ET AL. [2001] Initial sequencing and analysis of the human genome. Nature 409: 860-921.

[28] CARMICHAEL GG [2003] Antisense starts making more sense. Nat Biotechnol. 21: 371-372.

[29] MATTICK JS [2003] Challenging the dogma: The hidden layer of non-protein-coding RNAs in complex organisms. Bioessays. 25: 930-939.

[30] SOREK R, AST G AND GRAUR D [2002] Alu-containing exons are alternatively spliced. Genome Res 12: 1060-1067.

[31] DAGAN T, SOREK R, SHARON E. AST G AND GRAUR D [2004] AluGene: A database of Alu elements incorporated within protein-coding genes. Nucl Acids Res 32: D489-D492.

[32] ALUGENE. http://alugene.tau.ac.il/

[33] MCCLINTOCK B [1993] The Significance of responses of the genome to challenge. In: Lindsten J (ed.), Nobel Lectures, Physiology or Medicine 1981-1990, World Scientific Publishing Co., Singapore.

[34] KAZAZIAN HH JR [2004] Mobile elements: Drivers of genome evolution. Science 303: 1626-1632.

[35] WICHMAN HA, VAN DEN BUSSCHE RA, HAMILTON MJ AND BAKER RJ [1992] Transposable elements and the evolution of genome organization in mammals. Genetica 86: 287-293.

[36] TCHENIO T, CASELLA JF AND HEIDMANN T [2000] Members of the SRY family regulate the human LINE retrotransposons. Nucl Acids Res 28: 411-415.

[37] ROBINS DM AND SAMUELSON LC [1992] Retrotransposons and the evolution of mammalian gene expression. Genetica 86: 191-201.

[38] TOMILIN NV [1999] Control of genes by mammalian retroposons. Int Rev Cytol 186: 1-48.

[39] SPEEK M. [2001] Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. Mol Cell Biol 21: 1973-1985.

[40] NIGUMANN P, REDIK K, MATLIK K AND SPEEK M [2002] Many human genes are transcribed from the antisense promoter of L1 retrotransposon. Genomics 79: 628-634.

[41] LE GRICE SF [2003] In the beginning: Iitiation of minus strand DNA synthesis in retroviruses and LTR-containing retrotransposons. Biochemistry 42: 14349-14355.

[42] BIRO JC, BENYO B, SANSOM C, FORDOS G, MICSIK T AND BENYO Z [2003] A common periodic table of codons and amino acids. Biochem Biophys Res Com 306: 408-415.

[43] BIRO JC AND BIRO JMK [2004] Frequent Occurrence of recognition site-like sequences in the restriction endonucleases. BMC Bioinformatics 5: 30.

[44] PENNACCHIO LA AND RUBIN EM [2003] Apolipoprotein A5, a newly identified gene that affects plasma triglyceride levels in humans and mice. Arteriosclerosis Thrombosis Vascular Biol 23: 529-534.

[45] BRINKMAN RR, MEZEI MM, THEILMANN J, ALMQVIST E AND HAYDEN MR [1997] The likelihood of being affected with Huntington disease by a particular age, for a specific CAG size. Am J Hum Genet 60: 1202-1210.

[46] STEWART SA [2002] Multiple levels of telomerase regulation. Mol Interv 2: 481-483.

[47] SYLVESTRE P, COUTURE-TOSI E AND MOCK M [2003] Polymorphism in the collagen-like region of the Bacillus anthracis BclA protein leads to variation in exosporium filament length. J Bacteriol 185: 1555-1563.

RECEIVED 7-14-2004.

ACCEPTED 8-24-2004.