Novel conserved human DNA sequence supports expression of long genes

Jay C. Brown
Department of Microbiology, Immunology and Cancer Biology
University of Virginia School of Medicine
Charlottesville, Virginia, 22908

Corresponding author:
Dr. Jay C. Brown
Department of Microbiology, Immunology and Cancer Biology
University of Virginia School of Medicine Box 800734
Charlottesville, Virginia 22908
USA

Email: jcb2g@virginia.edu
Phone: 01-434-924-1814

Abstract

The study described here was carried out to explore the properties of a novel DNA sequence element found concentrated in five human chromosomes, 16-18, 20 and Y. The new DNA sequence, 1519 bp in length and called t1519, was found throughout the high abundance chromosomes, but only rarely elsewhere in the genome. BLAST analysis revealed the highest number of t1519 matches is in chromosome 18 (1429 matches) and the lowest in chromosome 16 (500). Nucleotide sequence homology was very good in all five high abundance human chromosomes; identities of 80% or more were observed in hundreds of annotated matches in each chromosome. Poor similarity, however, was observed in all non-human species examined. A few similarities were seen in chimpanzee chromosome 5 and in gorilla chromosome 7. Only poor sequence matches were found in mouse. Further analysis of human chromosomes 16 and 17 was performed with the goal of identifying the function of t1519. The focus was on t1519 matches found in protein coding genes and on the length and expression level of those genes. Results identified t1519 matches in both short (<200kb) and long (>200kb) genes while genes lacking t1519 were found only in the short population. Similarly, while the short genes included both high and low expression genes and t1519-positive and t1519-negative genes, all long genes were t1519 positive and weak in expression. The observations are interpreted to indicate that t1519 elements function to adjust the expression level of long genes so that the host can benefit from the functions they encode without permitting them to overextend the transcription machinery of the cell.

Keywords: human genome; long genes; transcription; novel DNA sequence element

Introduction

The study described here began with the unexpected observation of a novel DNA sequence element found in over 2000 copies in the human genome. The element, called t1519, is 1519 bp in length. It is found in high abundance in five human chromosomes, 16-18, 20 and Y, but only rarely in the remainder of the genome. This report begins with a description of the t1519 DNA sequence and proceeds to an inquiry of its function.

Studies of t1519 function suggest it has a role in expression of long genes, genes >200kb in length that together comprise ~10% of the human genome. Compared to other cell types, long genes are over-represented in neurons where they encode cell adhesion, receptor, and ion channel proteins [1,2]. Enrichment in the nervous system causes long genes to be involved in the etiology of neurological diseases including Rhett syndrome [3,4], Alzheimer's disease [5,6], and amyotrophic lateral sclerosis [7]. Other conditions such as ageing [8,9], and cancer [10,11] can also involve long genes. Long genes are present in the genomes of animals that evolved by the late Jurassic (150 mya), but are not found in descendants of animals that evolved earlier [1,12]. Analysis of long gene introns over evolutionary time supports the view that long genes evolved from shorter genes by elongation of the introns [1].

Properties of long gene structure indicate that their expression may be affected by mechanisms not found with short genes. If one assumes the accepted rate of transcription by RNA polymerase II (20-40 bp/sec [13,14]), then transcription of a long

gene would require hours (e.g., 2.1 hrs for a gene 300 kb in length) while short gene transcription would be accomplished in minutes. Depending on the number of RNA polymerase II molecules able to transcribe a single gene at once, therefore, the amount of mRNA synthesis from a long gene could be too small to meet the need for the cognate protein or so much that the level of long gene mRNA synthesis would affect the level of transcription machinery available for synthesis of other mRNAs. Additional levels of regulation may be required for long genes.

A relevant observation was contributed by King et al. [15] who studied long gene expression in cultures of mouse cortical neurons. Topoisomerase inhibitors (e.g., topotecan) were found to suppress long gene expression under conditions where the same effect was not observed with short genes. Since topoisomerases can unwind regions of overwound DNA produced during transcription activity [16-18], the result was interpreted to indicate that extra topoisomerase activity is required during transcription of long genes, and inhibition of this extra activity causes the observed suppression of long gene transcription.

Results

Properties of the t1519 sequence and its location in the human genome

t1519 was identified initially by comparing the DNA sequences of two human genes, CFHR3 and TSHR. The alignment yielded a match 1519 bp in length (Table 1) with a GC content of 39.5%. BLAST analysis beginning with the t1519 sequence revealed abundant matches in 5 of the 24 human chromosomes (Table 2). Hundreds of matches

were found in each of chromosomes 16-18, 20 and Y. The highest number was observed in chromosome 18 (1429) and the lowest in chromosome 16 (500). t1519 sequence matches were rare outside the five high abundance chromosomes. For instance, a group of 107 matches was found in the leftmost 5Mb of chromosome 6, and small numbers of matches were found in single genes elsewhere in the genome. Examples of the latter are: CSMD2 (chr1; 8 matches) and MYCBP2 (chr13; 4 matches).

Table 1: Nucleotide sequence of t1519, the novel human DNA repeated element described here

```
TTTTTACAAGGAAAATCTTTTTTATTTATTTATTGATTGATTTGTTATCATAATA
CTTTAAGTTTTAGGGCACATGTGCACAATGTGCAGGTTAGTTACATATGTATACA
TGTGCCATGCTGCTGCGCTGCACCCACTAACTCGTCATCTAGCATTAGGTATATC
TCCCAATGCTATCCCTCCCCCTTCCCCCCACCCCACAACAGTCCCCAGAGTGTGA
TGTTCCCCTTCCTGTGTCCATATGTTCTCATTGTTCAATTCCCACCTATGAGTGA
GAATATGCGGTGTTTGGTTTTTTGTTCTTGCGATAGTTTACTAAGAATGATGATT
TCCAATTTCATCCATGTCCCTACAAAGGACATGAACTCATCATTTTTTATGGCTG
CATAGTATTCCATGGTGTATATGTGCCACATTTTCTTAATCCAGTCTATCATTGT
TGGACATTTGGCTTGGTTCCAAGTCTTTGCTATTGTGAATAATGCCGCAATAAAC
ATACGTGTGCATGTGTCTTTATAGCAGCGTGATTTATAGTCCTTTGGGTATATAC
CCAGTAATGGGATGGCTGGGTCAAATGGTATTTCTAGTTCTAGATCCCTGAGGAA
TCGCCACACTGACTTCCACAATGGTTGAACTAGTTTACAGTCCCACCAACAGTGT
AAAAGTGTTCCTATTTCTCCACATCCTCTCCAGGACTTGTTGTTTCCTGACTTTT
TAATGATTGCCATTCTAACTGGTGTGAGATGGTATCTCATTGTGGTTTTGATTTG
CATTTCTCTGGTGGCCAGTGATGGTGAGCATTTTTTCATGTGTTTTTTGGCTGCA
TAAATGTCTTCTTTTGAGAAGTGTCTGTTCATGTCCTTCGCCCACTTTTTGATGG
GGTTGTTTGTTTTTTTCTTGTAAATTTGTTTGAGTTCATTGTAGATTCTGGATAT
TAGCCCTTTGTCAGATGACTAGGTTGCGAAAATTTTCTCCCATTTTGTAGGTTGC
CTGTTCACTCTGATGGTAGTTTCTTTTGCTGTGCAGAAGTTCTTTAGTTTAATTA
GATCCCATTTGTCAATTTTGGCTTTTGTTGCTATTGCTTTTGGTGTTTTAGACAT
GAAGTCCTTGCCCATGCCTATGTCCTGAATGGTAGTGCCTAGGTTTTCTTCTAGG
GTTTTTATGGTTTTAGGTCTAACGTTTAAGTCTTTAATCCATCTTGAATTGATTT
TTGTATAAGGTGTAAGGAAGGGATCCAGTTTCAGCTTTCTACATATGGCTAGCCC
GTTTTCCCAGCACCATTTATTAAATAGGGAATCCTTTCCTCATTGCTTGTTTTTC
TCAGGTTTGTCAAAGATCAGATAGTTGTAGATATGCGGCGTTATTTCTGAGGGCT
CTGTTCTGTTCCATTGATCTATGTCTCTGTTTTGGTGCCAGTACCATGCTGTTTT
```

GGTTACTGTAGCCTTGTAGTATAGTTTGAAGTCAGGTAGTGTGATGCCTCCAGCT
TTGTTCTTTTGGCTTAGGATTGACTTGGCAATAC

Table 2: Human chromosomes where t1519 is concentrated

chr	chr length (Mb)	t1519 matches	matches/Mb	genome
16	90.3	500	5.5	hg38
17	8.3	511	6.1	CHM13
18	80.4	1429	17.8	CHM13
20	64.4	589	9.1	CHM13
Y	57.2	655	11.5	hg38

The locations of t1519 sequence matches in the five high abundance chromosomes were downloaded from NCBI BLAST and plotted with Integrated Genome Viewer. The results showed that t1519 matches are found throughout the length of each high abundance chromosome (Fig. 1). Exceptions occur in the case of the Y chromosome where matches were found only in the leftmost 18Mb of the sequence, and in chromosome 16 where a large gap is found between positions ~30Mb and 50Mb. t1519 matches were absent from the centromere in each of the five chromosomes. A literature search was conducted to identify an earlier report of a sequence element with the properties of t1519, but none was found.

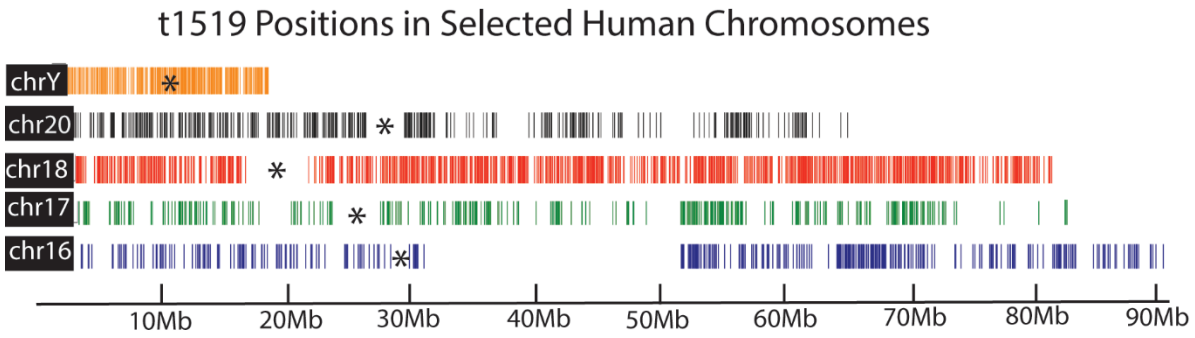

Fig. 1: Location of t1519 matches in the five high abundance human chromosomes. An asterisk indicates the location of the centromere of each chromosome. Note that t1519 matches are found only in the leftmost 18Mb of the Y chromosome. Data for the plots are shown in Supplementary Tables 1-5.

Sequence conservation of t1519

DNA sequence conservation was found to be very good in the five high abundance human chromosomes. In each there are 200 or more conserved regions 1121bp or greater in length and 87% or more in base pair identity (Table 3). In contrast, conservation was much weaker in three animal species examined (Table 4). Six regions of homology were observed in chimpanzee chromosome 5, all 748bp or greater in length with 97% bp identity. In gorilla chromosome 7, six homology regions were observed 285bp or more in length and 96% identity. Homology was much lower in mouse where one weak t1519 match was found in each of 3 different chromosomes.

Table 3: Conservation of t1519 sequence matches in high abundance human chromosomes

		Nucleotide		
Chr	**Match no.**	**matches/total**	**%**	**Score**
16	1	1452/1495	97	2521
17	1	1465/1503	97	2564
18	1	1441/1477	98	2527
20	1	1436/1471	98	2521
Y	1	1465/1499	98	2579
16	100	1393/1497	93	2183
17	100	1185/1214	98	2082
18	100	1414/1469	96	2407
20	100	1404/1494	94	2255
Y	100	1383/1503	92	2109

16	200	1061/1121	95	1736
17	200	1192/1372	87	1520
18	200	1369/1465	93	2172
20	200	1335/1486	90	1901
Y	200	1320/1504	88	1742

Table 4: Comparison of t1519 conservation in human and selected animal species

			Nucleotide		
Species	**Chr**	**Match no.**	**matches/total**	**%**	**Score**
human	17	1	1465/1503	97	2564
human	17	3	1427/1461	98	2510
human	17	6	1446/1493	97	2494
chimpanzee	5	1	1456/1501	97	2519
chimpanzee	5	3	1011/1047	97	1753
chimpanzee	5	6	671/748	97	950
gorilla	7	1	1452/1498	97	2510
gorilla	7	3	1282/1489	86	1580
gorilla	7	6	273/285	96	460
mouse	12	1	359/484	74	176
mouse	4	1	162/211	77	115
mouse	15	1	178/239	74	104

t1519 function

Further studies of t1519 were performed to attempt to define its function. This effort was focused on t1519 elements present in the body of protein coding genes as these seemed to be over-represented in the overall population. For example, among the 500 t1519 elements located in chromosome 16, 200 (40%) were found in protein coding genes (Table 5). In chromosome 17 the percentage was 45%. t1519 elements not located in protein coding genes were found in intergenic regions, genes specifying non-coding RNAs and pseudogenes.

Table 5: Proportion of t1519 sequences inside protein coding genes of human chromosomes 16 and 17

t1519 location	Chr16	Chr17
Inside a protein coding gene	200 (40%)	228 (45%)
Outside any protein coding gene	300 (60%)	283 (55%)
Total	500 (100%)	511 (100%)

The strategy for clarifying t1519 function was to begin with a population of protein coding genes from chromosome 16 and compare those that contain t1519 matches with those that do not. The properties compared were gene length and expression level of the gene. The same analysis was performed with protein coding genes from chromosome 17. Gene length and expression level (GTEx-RNA-seq) values were downloaded from UCSC Genome Browser.

Analysis was performed with 158 genes from chromosome 16, 79 genes with t1519 and 79 without. The count of each gene type was plotted against gene length (Fig. 2a). The results showed that while all genes lacking t1519 have lengths of <150kb, all 20 genes >150kb in length have at least one t1519 match. A similar result was obtained with chromosome 17. The chromosome 17 analysis was carried out with a total of 119 genes, 61 with and 58 without t1519, respectively. The results showed that while all genes lacking t1519 were <175kb in length, all chromosome 17 genes >175kb in length were t1519 positive (Fig. 2b). The observations with chromosome 16 and 17 genes are interpreted to indicate that t1519 has a function that relates to some aspect of long genes that is different from that in short genes.

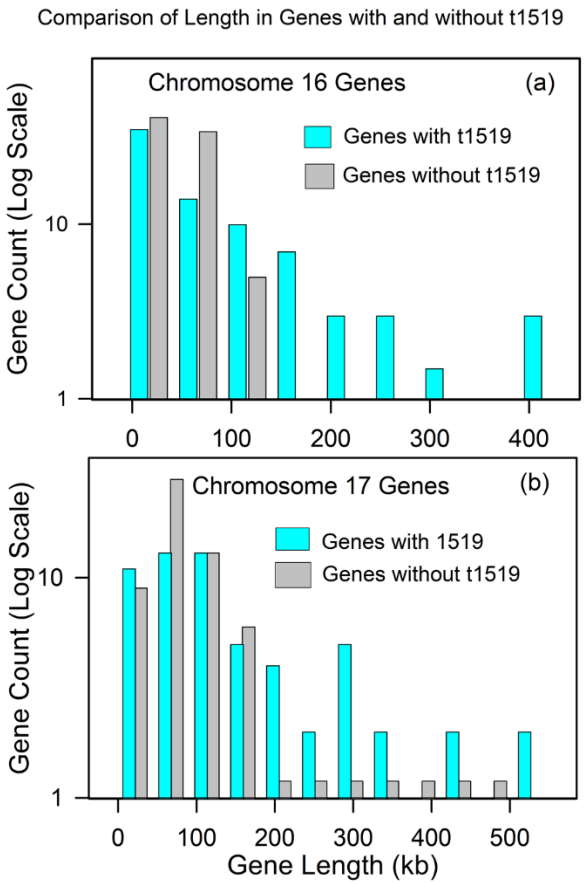

Fig. 2: Plot comparing the length of t1519 positive with t1519 negative genes. Gene samples were derived from chromosome 16 (a) and 17 (b). Note that in both cases all long genes (>200kb) are t1519 positive.

Studies of gene expression were performed with the same populations of chromosome 16 and 17 genes employed for the analysis of gene length described above. In the case of expression studies, gene expression was plotted against gene length in t1519 positive and negative populations (Fig. 3). The results with chromosome 16 genes demonstrated that while t1519 negative genes could have either high (>600 TPM) or low (<200 TPM) expression, t1519 positive genes all have low (<100 TPM) expression (Fig 3a). Expression of a t1519 positive gene was low whether its length was low or

high. The results with chromosome 17 genes resembled those of chromosome 16 except that some short, t1519 positive genes were found to have high expression. However, if a gene was long and t1519 positive, then its expression level was always low (<60 TPM). The results are interpreted to indicate that there is an effect of t1519 on gene expression that is specific for long genes. Short, t1519 positive genes can be either high or low in expression, but long, t1519 positive genes all have low expression.

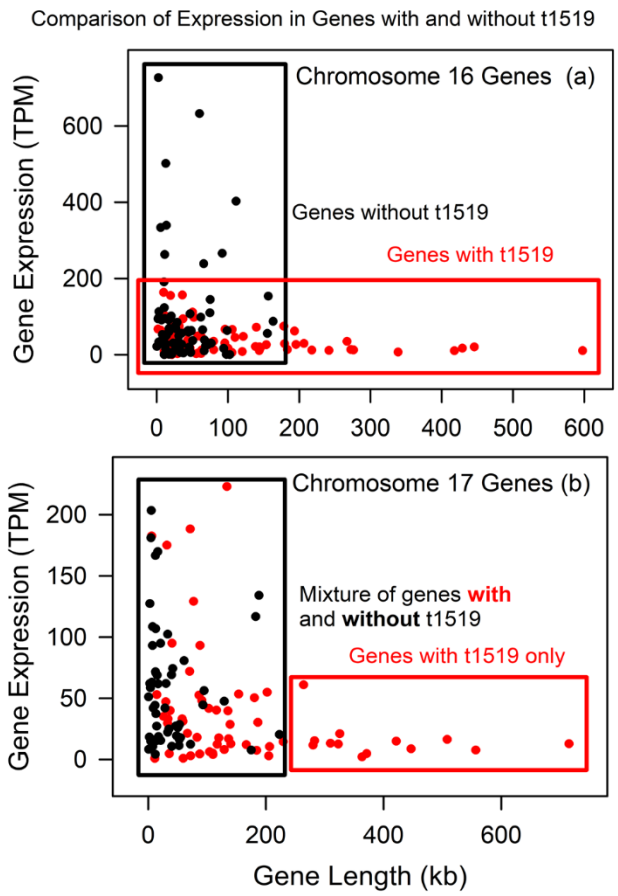

Fig. 3: Plot showing the expression level as a function of gene length in chromosome 16 t1519 positive (red) and t1519 negative (black) genes (a). Panel (b) shows the same information from chromosome 17 genes. Boxes are for clarity only. They do not contain any new information. Note that all long, t1519 positive genes are low in expression.

Discussion

The interpretations offered above for the function of t1519 are consistent with what is known about transcription of long genes. Like the processes of DNA replication and recombination, transcription from a DNA template involves overwinding of the DNA [17,18]. For genes less than a certain length, the amount of overwinding can be accommodated by cellular topoisomerases or other processes. For longer genes, however, topoisomerases need to be deployed to relax the DNA or transcription stalls [15,16]. It is suggested that t1519 sequences are involved in providing the topoisomerase or similar unwinding activity required for transcription of long genes. The proposed role for t1519 sequences is compatible with its presence throughout the length of the high abundance human chromosomes (i.e., chromosomes 16-18, 20 and Y), with its prevalence in protein coding genes and with its presence in long genes. The absence of t1519 sequences from many human chromosomes continues to require explanation.

The results shown in Fig. 3 suggest t1519 has a second activity or consequence related to its ability to enable expression of long genes. Expression of all long genes is found to be very low in both chromosome 16 and 17 populations, and all the genes are t1519 positive. This observation indicates that while t1519 can enable long gene transcription, it also restricts expression to a low level. This combination of properties provides the host animal with the extra genetic information present in a long gene without the risk of absorbing too much of the cell's RNA synthesis capacity with long gene transcription.

The results shown in Figs. 2 and 3 suggest that t1519 elements may have functions beyond those involving long genes. Short genes are an example. Many short genes contain t1519 elements in both the chromosome 16 and 17 populations. It may be, therefore, that some short genes require template DNA unwinding as long genes do and others do not. Alternatively, the effect of t1519 to suppress gene expression may benefit some short genes as well as long ones.

A striking feature of the results described here is the high degree of DNA sequence conservation in human t1519 elements. Hundreds of high-quality sequence matches were observed in each of the high abundance chromosomes suggesting that a high degree of sequence identity is required for t1519 functions. An equally noteworthy feature is the poor conservation in the t1519 sequences in the animal species examined. (i.e., chimpanzee, gorilla, and mouse; Table 4). This observation indicates that t1519 elements may be an important part of what makes humans distinct from even our closest animal relatives. For instance, the extra information present in human long genes compared to chimpanzee, gorilla long genes may be important for human-specific functions. The presence of highly human-specific t1519 DNA sequence information also suggests that human evolution involving t1519 elements may be actively occurring at the present time.

Materials and Methods

Pairwise DNA sequence alignments were determined with the local alignment tool Matcher available from the EMBL-EBI Pairwise Sequence Alignment suite (https://ebi.ac.uk/Tools/psa). Other sequence alignments were determined with NCBI BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi) and rendered graphically with Integrated Genome Viewer (https://igv.org). Gene lengths and expression levels were downloaded from UCSC Genome Browser (https://genome.ucsc.edu) and analyzed with Excel and SigmaPlot 14.5.

Acknowledgements

I gratefully acknowledge the assistance of Anahita Golchin in the early stages of this investigation.

References

1. McCoy MJ, Fire AZ. Intron and gene size expansion during nervous system evolution. BMC Genomics. 2020;21(1):360.

2. Brown JC. Role of Gene Length in Control of Human Gene Expression: Chromosome-Specific and Tissue-Specific Effects. International Journal of Genomics. 2021;2021:8902428.

3. Chahrour M, Zoghbi HY. The story of Rett syndrome: from clinic to neurobiology. Neuron. 2007;56(3):422-37.

4. Gabel HW, Kinde B, Stroud H, Gilbert CS, Harmin DA, Kastan NR, et al. Disruption of DNA-methylation-dependent long gene repression in Rett syndrome. Nature. 2015;522(7554):89-93.

5. Barbash S, Sakmar TP. Length-dependent gene misexpression is associated with Alzheimer's disease progression. Sci Rep. 2017;7(1):190.

6. Soheili-Nezhad S, van der Linden RJ, Olde Rikkert M, Sprooten E, Poelmans G. Long genes are more frequently affected by somatic mutations and show reduced

expression in Alzheimer's disease: Implications for disease etiology. Alzheimers Dement. 2021;17(3):489-99.

7. Lagier-Tourenne C, Polymenidou M, Hutt KR, Vu AQ, Baughn M, Huelga SC, et al. Divergent roles of ALS-linked proteins FUS/TLS and TDP-43 intersect in processing long pre-mRNAs. Nat Neurosci. 2012;15(11):1488-97.

8. Stoeger T, Grant RA, McQuattie-Pimentel AC, Anekalla KR, Liu SS, Tejedor-Navarro H, et al. Aging is associated with a systemic length-associated transcriptome imbalance. Nat Aging. 2022;2(12):1191-206.

9. Ibanez-Sole O, Barrio I, Izeta A. Age or lifestyle-induced accumulation of genotoxicity is associated with a length-dependent decrease in gene expression. iScience. 2023;26(4):106368.

10. Smith DI, McAvoy S, Zhu Y, Perez DS. Large common fragile site genes and cancer. Semin Cancer Biol. 2007;17(1):31-41.

11. Sahakyan AB, Balasubramanian S. Long genes and genes with multiple splice variants are enriched in pathways linked to cancer and other multigenic diseases. BMC Genomics. 2016;17:225.

12. McCoy MJ, Fire AZ. Ancient origins of complex neuronal genes. bioRxiv. 2023.

13. Ardehali MB, Lis JT. Tracking rates of transcription and splicing in vivo. Nat Struct Mol Biol. 2009;16(11):1123-4.

14. Core LJ, Waterfall JJ, Lis JT. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science. 2008;322(5909):1845-8.

15. King IF, Yandava CN, Mabb AM, Hsiao JS, Huang HS, Pearson BL, et al. Topoisomerases facilitate transcription of long genes linked to autism. Nature. 2013;501(7465):58-62.

16. Mabb AM, Kullmann PH, Twomey MA, Miriyala J, Philpot BD, Zylka MJ. Topoisomerase 1 inhibition reversibly impairs synaptic function. Proc Natl Acad Sci U S A. 2014;111(48):17290-5.

17. Vos SM, Tretter EM, Schmidt BH, Berger JM. All tangled up: how cells direct, manage and exploit topoisomerase function. Nat Rev Mol Cell Biol. 2011;12(12):827-41.

18. Baranello L, Levens D, Gupta A, Kouzine F. The importance of being supercoiled: how DNA mechanics regulate dynamic processes. Biochim Biophys Acta. 2012;1819(7):632-8.

Supplementary Tables (.bed files)

Supplementary Table 1: t1519 locations in human chromosome 16

https://med.virginia.edu/brown/wp-content/uploads/sites/506/2023/11/chr16_t1519.bed_.txt

Supplementary Table 2: t1519 locations in human chromosome 17

https://med.virginia.edu/brown/wp-content/uploads/sites/506/2023/11/chr17_t1519.bed_.txt

Supplementary Table 3: t1519 locations in human chromosome 18

https://med.virginia.edu/brown/wp-content/uploads/sites/506/2023/11/chr18_t1519.bed_.txt

Supplementary Table 4: t1519 locations in human chromosome 20

https://med.virginia.edu/brown/wp-content/uploads/sites/506/2023/11/chr20_t1519.bed_.txt

Supplementary Table 5: t1519 locations in human Y chromosome

https://med.virginia.edu/brown/wp-content/uploads/sites/506/2023/11/chrY_t1519.bed_.txt