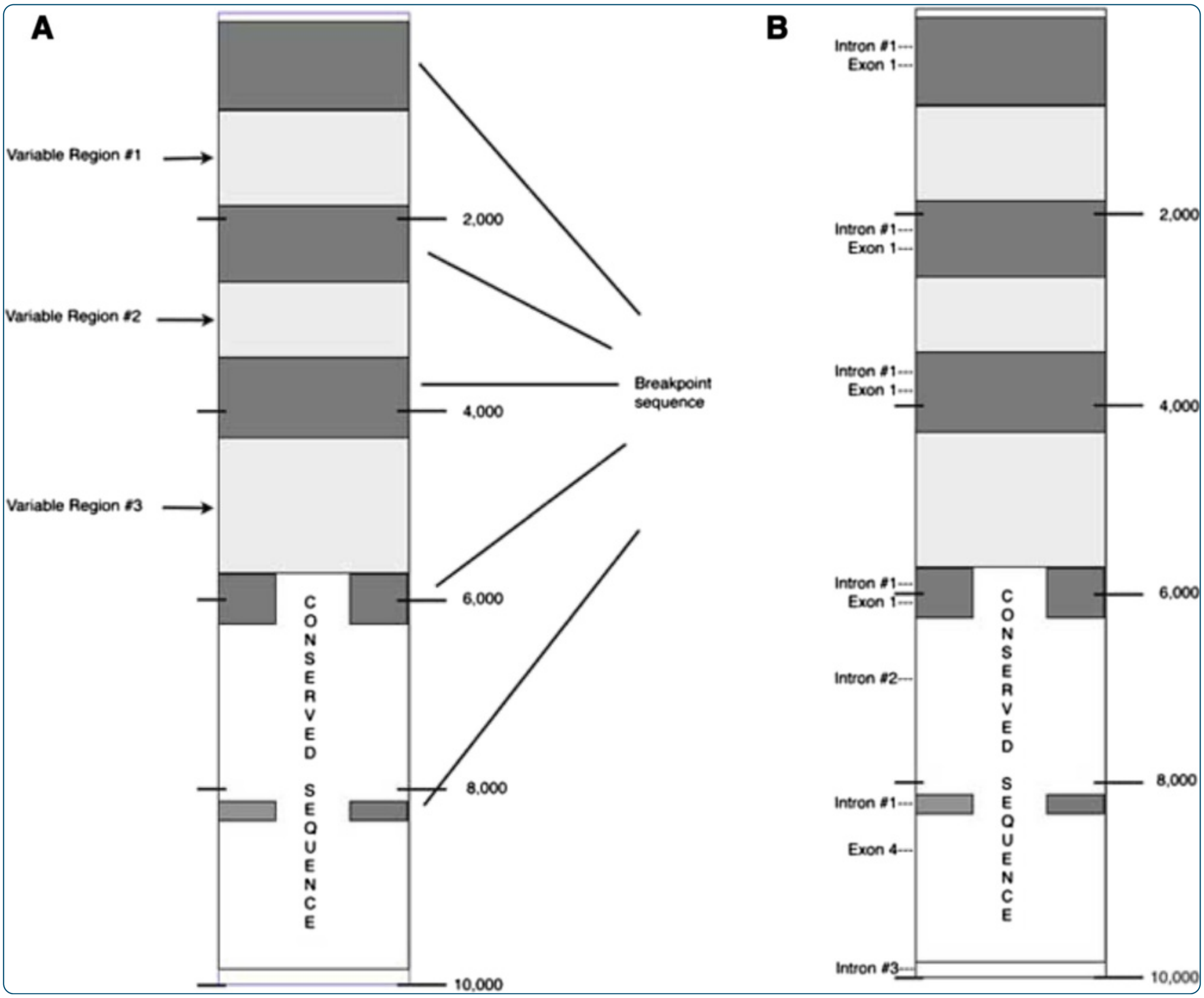
BMC Genomics



Complexity of a small non-protein coding sequence in chromosomal region 22q11.2: presence of specialized DNA secondary structures and RNA exon/intron motifs

Delihas

BioMed Central

Delihas *BMC Genomics* (2015) 16:785
DOI 10.1186/s12864-015-1958-6

Delihas *BMC Genomics* (2015) 16:785
DOI 10.1186/s12864-015-1958-6

RESEARCH ARTICLE Open Access

Complexity of a small non-protein coding sequence in chromosomal region 22q11.2: presence of specialized DNA secondary structures and RNA exon/intron motifs

Nicholas Delihas

Abstract

Background: DiGeorge Syndrome is a genetic abnormality involving ~3 Mb deletion in human chromosome 22, termed 22q.11.2. To better understand the non-coding regions of 22q.11.2, a small 10,000 bp non-protein-coding sequence close to the DiGeorge Critical Region 6 gene (*DGCR6*) was chosen for analysis and functional entities as the homologous sequence in the chimpanzee genome could be aligned and used for comparisons.

Methods: The GenBank database provided genomic sequences. In silico computer programs were used to find homologous DNA sequences in human and chimpanzee genomes, generate random sequences, determine DNA sequence alignments, sequence comparisons and nucleotide repeat copies, and to predicted DNA secondary structures.

Results: At its 5′ half, the 10,000 bp sequence has three distinct sections that represent phylogenetically variable sequences. These Variable Regions contain biased mutations with a very high A + T content, multiple copies of the motif TATAATATA and sequences that fold into long A:T-base-paired stem loops. The 3′ half of the 10,000 bp unit, highly conserved between human and chimpanzee, has sequences representing exons of lncRNA genes and segments of introns of protein genes. Central to the 10,000 bp unit are the multiple copies of a sequence that originates from the flanking 5′ end of the translocation breakpoint Type A sequence. This breakpoint flanking sequence carries the exon and intron motifs. The breakpoint Type A sequence seems to be a major player in the proliferation of these RNA motifs, as well as the proliferation of Variable Regions in the 10,000 bp segment and other regions within 22q.11.2.

Conclusions: The data indicate that a non-coding region of the chromosome may be reserved for highly biased mutations that lead to formation of specialized sequences and DNA secondary structures. On the other hand, the highly conserved nucleotide sequence of the non-coding region may form storage sites for RNA motifs.

Keywords: Chromosome region 22q.11.2, Translocation breakpoint sequences, Biased mutations, DNA secondary structure, RNA exons, Introns

Background

DiGeorge Syndrome is part of a group of genetic disorders that occur in humans termed 22q11.2 deletion syndrome [1]. The DiGeorge disorder involves ~3 Mb deletion of the chromosomal 22q11.2 region that results in loss of approximately ~40 protein genes. Clinical manifestations of a 22q11.2 deletion are pleiotropic and can include congenital heart disease, developmental problems, neurological abnormalities, immune system abnormalities, learning disabilities and psychiatric problems that also includes autism [1–4].

Deletions in chromosomal region 22q11.2 can occur from an unequal crossover between low copy repeats (LCR)s present in this region of chromosome 22. LCRs can contain very long A + T-rich stem loops that are termed palindromic A + T-rich repeats (PATRR) [5–11]. These have been cloned and sequenced and appear to be central to translocation [12]. PATRRs can form cruciforms and are prone to breakage that can lead to

Correspondence: Nicholas.delihas@stonybrook.edu
Department of Molecular Genetics and Microbiology, School of Medicine, Stony, Brook University, Stony Brook, NY 11794, USA

chromosomal translocation. The stem loop sequences are also referred to as translocation breakpoint hot spots. Some of these have flanking sequences that are conserved or partially conserved between closely related PATRR elements.

To better understand the nature of non-coding genomic regions in a 22q.11.2 deletion, an analysis was made of a 10,000 bp non-protein coding region (GRCh38 primary assembly, coordinates:18890337–18900336), which is just upstream of the DiGeorge Critical Region 6 gene (*DGCR6*). *DGCR6* is homologous to a *Drosophila melanogaster* gene, whose protein product may participate in gonadal cell development and other cell functions [13]. The 10,000 bp region, albeit representing a small fraction of the 3 Mb region, was chosen for analysis as the homologous chimpanzee sequence for the most part has been determined. Importantly, the *DGCR6* gene has been annotated in both species and this provided the guide post for alignment of sequences between species. In addition, alignment was possible even with the presence of redundant sequences (low-complexity sequences), and thus, comparisons that reveal high sequence identity and/or variation could be made. Central to this region and other regions of 22q11.2 is the presence of multiple copies of segments of the Translocation Breakpoint Type A sequence (GenBank, AB261997) [12]. We find numerous segments of translocation breakpoint flanking sequence insertions within the 10,000 bp region, but these lack the breakpoint hot spot stem loop sequences (PATRRs). Several of the breakpoint flanking sequences are not present in the homologous chimpanzee region. This enabled us to delineate three sequence Variable Regions, #1-#3 and these are in the 5′ half of the 10,000 bp unit.

The three phylogenetically Variable Regions are very A + T-rich, in both humans and chimpanzee. These regions mutate very rapidly, however, in humans, mutations significantly accelerate the formation of a nine bp sequence, TATAATATA [9] and there also are sequences that fold into long A:T-rich stem loops, which are better formed and longer in humans relative to the chimpanzee. Of importance, the 22q11.2 region is one of the most susceptible regions of the human genome to undergo genetic recombination [12], although the 10,000 bp region is not in the region observed to undergo repeated translocation, which is the LCR-B chromosomal region [10].

To add to the complexity of the 10,000 bp region, we find that translocation breakpoint flanking sequences are carriers of a specific exon sequence that is present in long non-coding RNAs (lncRNA)s and in addition, segments of mRNA intron sequences; these are abundantly found in the 10,000 bp unit. The motifs may represent reservoirs for use in possible formation of new lncRNA and protein genes.

This paper presents new and multifaceted concepts in genomics, which include the spread of RNA exon and intron motifs within the genome by duplication of translocation breakpoint sequences, genomic regions reserved for translocation breakpoint sequences that may participate in formation of highly biased mutations in A + T-rich regions, and phylogenetically stable storage regions in the genome that contain both lncRNA exon and mRNA intron sequences.

Results

The 3′ end of the 10,000 bp non-coding region is situated 6039 bp upstream of the *DGCR6* gene in Homo sapiens chromosome 22 (GRCh38 Primary Assembly, coordinates:18890337–18900336). This region has blocks of repeats and redundant sequences that are of low complexity and these create problems in sequence alignment. However, alignment in the 10,000 bp region was possible because of the presence of translocation breakpoint sequence inserts that serve as guideposts. Alignment of sequences with the analogous chimpanzee sequence, which is 5892 bp upstream of the *DGCR6* isoform 1 gene (Pan troglodytes chromosome 22, Pan_troglodytes-2.1.4, coordinates: 17300774 to 17307562) is shown in Additional file 1: Figure S1, and a schematic of the 10,000 bp region is in Fig. 1. Three global alignment programs [14–16] were employed to verify the accuracy of the overall alignment. “Edge effects” at A + T-rich redundant regions occur with different alignment programs but this had a negligible effect on the overall alignment pattern as the translocation breakpoint flanking sequence inserts, which are present in the human sequence but are mostly missing in the chimpanzee region (see Additional file 1: Figure S1), sufficiently delineate the variable sequence regions, serve as guideposts and allow for Variable Region analyses.

Alignment at the 5′ half of the human 10,000 bp segment shows three significant variable sequence blocks, termed Variable Regions #1- #3 (Fig. 1a). The inserts are breakpoint Type A flanking sequences, and these are a major component of the 5′ half of the human 10,000 bp unit. However, the nucleotide sequence lengths within each Variable Region differ when the human sequence is compared with the homologous chimpanzee sequence. This is not due to breakpoint sequence additions, but due to other added sequences that are only present in the human segments of the Variable Regions. There are no additions in the comparable chimpanzee regions, e.g., see human positions 911–1953, Additional file 1: Figure S1.

On the other hand, the 3′ half of the 10,000 segment shows a very high nucleotide sequence identity (97 %) between human and chimpanzee sequences, however, there are two partial breakpoint flanking sequences

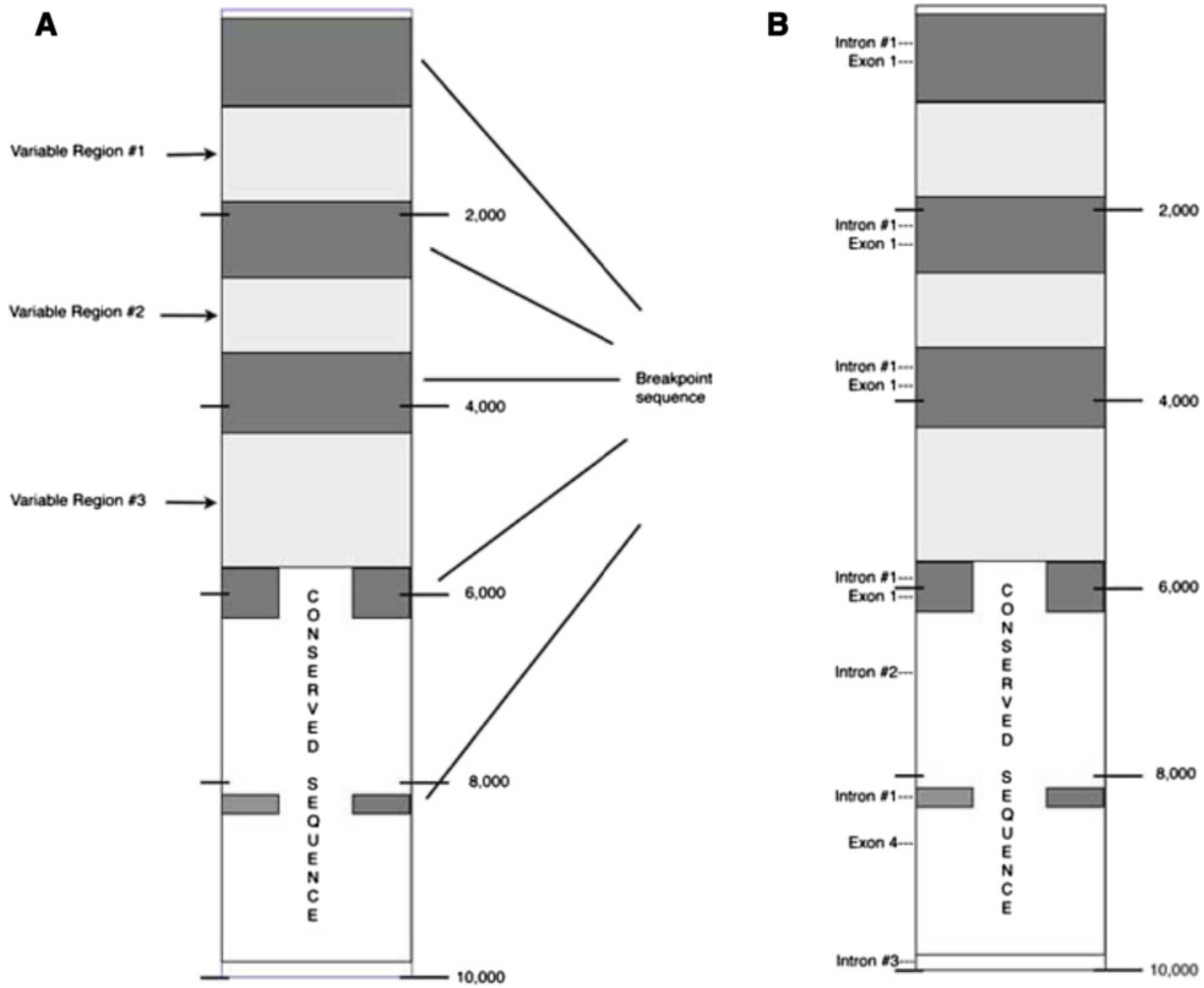


Fig. 1 Schematic of the human 10,000 bp region upstream of *DGCR6*. **a** Locations of breakpoint flanking sequences. **b** Locations of exon 1 and exon 4 and introns #1-#3

present in both the human and chimpanzee regions and they are also conserved phylogenetically. Figure 1b shows the locations of intron and exon sequences that are present in the 10,000 bp unit.

Translocation breakpoint sequences and secondary structures

Collaborations between Japanese and American investigators resulted in pioneer work on the characterization of translocation breakpoint hot spot secondary structures and functions of these structures in genetic exchange [10, 17, 18]. In addition, another group, by using biophysical calculations has shown that translocation frequency is very closely related to stem loop ability to form DNA cruciform structures [19]. One palindromic sequence found on chromosomal segment 22q11.2 is Type A (NCBI GenBank: AB261997.1). Two variations, Types B and C are also known. They have minor sequence changes, however Type A has a repeat of the first 363 bp of the 5′ end sequence at its 3′ end sequence. In addition, breakpoint sequences present on chromosome 11 have also been well-characterized [7]. PATRR breakpoint hot spot sequences fold into very long stem-loops [7, 10, 20, 21]. A total of twelve PATRR sequences and their translocation frequencies have been described [10]. We analyzed secondary structures of the twelve PATRR sequences, however, two PATRRs that exhibit extreme examples of translocation frequency are described here. These two differ by over a factor of 100 in translocation frequency [10]. The Chr 22 TYPE C (accession number AB538237.2) sequence shows a near perfect 294 bp stem, two TATAATATA motifs on the stem situated close to the top apex loop, but has an internal bulge with 3 nt on both sides of the stem (Additional file 2: Figure S2). The PATRR structure TYPE C from Chr 22 exhibits one of the highest translocation frequencies [10]. In contrast, a PATRR sequence from Chr 11 (accession number AF391128) is one of four PATRRS that exhibits a very low frequency of translocation [10]. Its predicted secondary structure shows a more imperfect stem with two large looped out regions, has a much smaller stem (87 bp), but it does have a TATAATATA motif close to the top apex loop on the 3′ side (Additional file 3: Figure S3).

A comparison of secondary structures and translocation frequencies of the twelve PATRR sequences suggests the following for PATRRS that display a high frequency of translocation. A near perfect long stem consisting of ~200 bp or more, a small top apex loop (<5 nt), greater than 90 % A:T base pairs in the upper third of the stem and a moderate abundance (~40 %) of G:C pairs in the bottom 2/3 of the stem appear to be important. Small internal loops in the stem are tolerated, but the presence of large internal loops, protruding stem loops, or short stems do not appear be to conducive to high frequency translocation. The TATAATATA sequence close to the top apex loop is common to most PATRRs. However, the motif is not found in all PATRR structures that exhibit translocation, e.g. the PATRR stem loop in NCBI Gene Bank accession #AB235190 nt sequences, albeit this example displays a low frequency of translocation [7].

The translocation breakpoint Type A, in addition to carrying the A + T palindromic breakpoint sequence, surprisingly contains two unrelated motifs; these reside on the 5′ flanking side of the breakpoint hot spot sequence (Additional file 4: Figure S4). These are RNA motifs that consist of an exon sequence found as exon 1 in different lncRNA transcripts with a high sequence identity, and a partial sequence of an intron found in different mRNA transcripts, also with a high identity.

10,000 bp Unit Variable Regions

An analysis was made of the three Variable Regions found in the 5′ half of the human 10,000 bp unit that are very rich in A + T residues. In addition, the number of copies of the 9 nt sequence TATAATATA [12] was determined in all three Variable Regions. Results show that there are multiple copies of the TATAATATA motif present in each of the three Variable Regions, but humans contain a significantly larger number than present in the chimpanzee (Table 1). There appears to be a marked bias towards adding and/or maintaining the TATAATATA motif in Variable Regions in both human and chimpanzee genomes. For example, human Region #1 contains 94 % A + T residues with a length of 1054 nt. Twenty-five random sequence samples with the same parameters (A + T percentage and length) show that on the average, 0.8 copies of TATAATATA/random sample, whereas Variable Region #1 in humans has 38 and the chimpanzee has 20 (Table 1). The p-value for the human TATAATATA copy number (38) vs the copy number from random sequences (0.8) has been determined by a conservative nonparametric test, Wilcoxon's signed rank test [22]. The p-value is <0.0001. Thus, the human sequence has an almost a 50-fold greater number of the conserved 9 bp sequence relative the 25 random sequence samples. However, the Variable Regions also maintain a high copy number of a few very closely related sequences such as TATTATATA as well (data not shown); thus, a bias extends to closely related sequences as well.

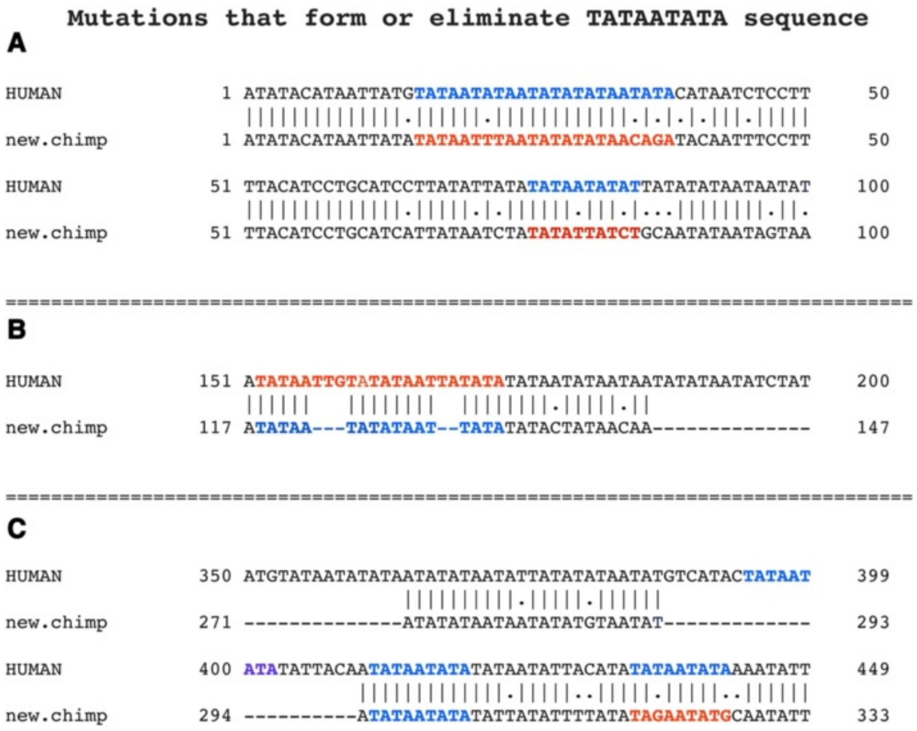
Table 1 TATAATATA copies in DNA variable regions

Variable Region	Human coordinates	TATAATATA copies Human	chimpanzee	Ratio[a]
1.	911-1964	38	20	1.9
2.	2753-3484	20	8	2.5
3.	4365-5760	26	23	1.1

[a] Ratio of TATAATATA copies, human/chimpanzee

A comparison of aligned sequences in the Variable Regions shows that the human sequence has expanded greatly compared to the chimpanzee; however, the additional sequences are not related to translocation breakpoint flanking sequences. Analyses of alignments show that both point mutations and the additional sequences in the human genome contribute to the greater number of TATAATATA copies in the human sequence. Figure 2 shows examples of bp mutations and/or addition of nt sequences that create as well as destroy the TATAATATA motif in human sequences compared to that of the chimpanzee, as well as a conservation of the motif between the two species.

Variable regions were also analyzed for DNA secondary structure features. The structure of Chr22 TYPE C Accession:AB538237.2 PATRR (Fig. 3a), which displays a very high frequency of translocation [10] is used here as a model for DNA secondary structure and high translocation. In addition, the predicted structure for a typical sample random sequence is also shown (Fig. 3b). All three Variable Regions of the human 10,000 bp segment show at least one long A:T base pair-rich stem loop structure, albeit the human Variable Region #1 sequence folds into a poorly formed stem loop. We use the sequence from Variable Region #3 as a model, which displays the best-formed stem loops. Human and chimpanzee predicted secondary structures from this region are shown in Fig. 4a and B, respectively. The human structure shows two long stem loops that are fairly well formed; the chimpanzee has one. The lengths of the



```
           Mutations that form or eliminate TATAATATA sequence
A

HUMAN                1 ATATACATAATTATGTATAATATAATATATATAATATACATAATCTCCTT     50
                       ||||||||||||||.||||||.||||||||||||.|.|.|.|||.|||||
new.chimp            1 ATATACATAATTATATATAATTTAATATATATAACAGATACAATTTCCTT     50

HUMAN               51 TTACATCCTGCATCCTTATATTATATATAATATATTATATATAATAATAT    100
                       ||||||||||||||.|||||.|.||||||.|||.|...||||||||.||.
new.chimp           51 TTACATCCTGCATCATTATAATCTATATATTATCTGCAATATAATAGTAA    100

=============================================================================
B

HUMAN              151 ATATAATTGTATATAATTATATATATAATATAATAATATATAATATCTAT    200
                       ||||||   ||||||||  ||||||||.|||||.||
new.chimp          117 ATATAA---TATATAAT--TATATATACTATAACAA--------------    147

=============================================================================
C

HUMAN              350 ATGTATAATATATAATATATAATATTATATATAATATGTCATACTATAAT    399
                                     ||||||||||.|||||.||||||
new.chimp          271 --------------ATATATAATAATATATGTAATAT-------------    293

HUMAN              400 ATATATTACAATATAATATATATAATATTACATATATAATATAAAATATT    449
                                 |||||||||||||.|||||..|||||.|||||..||||||
new.chimp          294 ----------ATATAATATATATTATATTTTATATAGAATATGCAATATT    333
```

Fig. 2 Alignment of human and chimpanzee sequences that shows formation or loss of TATAATATA motif. Three separate alignment programs gave similar alignments in the small genomic segments shown above. The nature of mutations, whether a base change, sequence addition or oligo-insertion in the human sequence occurred is shown. Color code: blue, TATAATATA motif formed, red, mutations destroy motif. **a** bp substitution creates TATAATATA motif in human sequence (however, we can not rule out a point mutation in the chimpanzee sequence destroyed a TATAATATA motif). **b** Oligo-bp insertions in human sequence eliminates two overlapping TATAATATA motifs. **c** Oligo-bp insertion in human sequence adds motif, another motif is conserved between two species, and two bp substitutions create a motif in human. Alignments are by Emboss Needle (www.ebi.ac.uk/Tools/psa/emboss_needle/)

human stem loops are 120 bp (stem loop 2) and 109 bp (stem loop 1); the chimpanzee structure shows 106 bp (Table 2). A comparison of the long stem loops between human and chimpanzee shows that the human structures have fewer looped out regions and smaller stem protruding "mini-stem loops" (Fig. 4). The entire Variable Region #3 is also more thermodynamically stable in humans than in chimpanzee, 257 kcal/mol vs 164 kcal/mol, respectively (Table 3) but the Gibbs free energies for the stem loops alone are only moderately more stable in human samples compared to the chimpanzee. Overall, the human stem loop structures are much closer to the model translocation hot spot structure (Fig. 3a) than that of the chimpanzee.

The apex "hairpin loop" structure (Fig. 5) is also better formed in humans compared to that of the chimpanzee with 4 bases comprising the single stranded loop for the human loop but 18 for the chimpanzee loop (Fig. 5b, c), but both human (stem loop 2) and the chimpanzee have TATAATATA motifs on the 5′ and 3′ sides of the loop, and the chimpanzee stem loop has 3 copies, two of which overlap on the 5′ side. The model translocation breakpoint Type C stem loop has 5 bases in the apex loop and TATAATATA motifs on both 5′ and 3′ sides of the loop (Fig. 5a)

A striking difference between Variable Region #3 stem loops and the breakpoint Type C stem loop is the sharp contrast in stem loop delta G values (Table 2). This is due to the greater number of bp and a much greater number of G:C pairs present in the Type C hairpin stem. Translocation Type C stem has ~5 times as many G:C pairs/bp as human stem loop 2 (Table 4). The chimpanzee stem has about a tenth of Type C stem. G:C pairs may be crucial to maintaining the stem, as its upper apex portion may extensively breathe or unfold since it is very A:T base pair-rich.

Although approaching a breakpoint hot spot stem loop (Type C) structure, the human Variable Region #3 secondary structure lacks important signatures of Type C and other PATRR stem loops. Thus, a further maturation is needed to form a structure that is more analogous to a PATRR.

Random sequences of the same length and A + T content as the human Variable Region #3 stem loop 2 fold into imperfect stem loops with significant

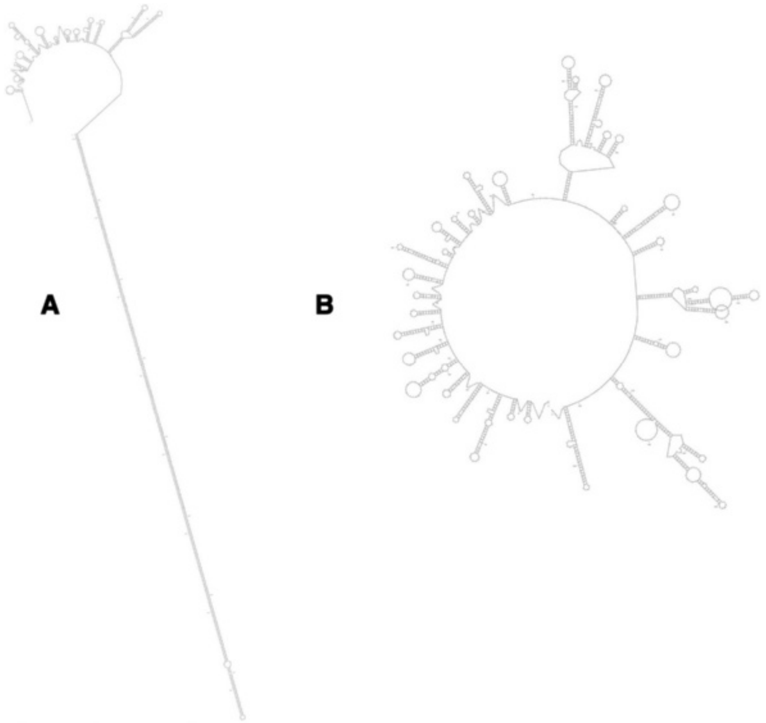


Fig. 3 Predicted DNA secondary structures generated by mfold (http://mfold.rna.albany.edu/?q=mfold/dna-folding-form). **a** Type C PATRR. Sequence from Accession:AB538237.2. **b** Random sequence determined by use of Molbiol.ru: (http://molbiol.ru/eng/scripts/01_16.html) and generated from 1393 nucleotides with 8.3 % G + C

numbers of "mini-stem loops" protruding from the main stem (e.g., see Fig. 3b). A total of forty random sequences were generated and analyzed for secondary structure. They display widely different types of structures, but importantly, seven out of the forty random sequences display very long stem loops (Additional file 5: Table S1). On the other hand, these structures show many imperfections in stems with bulged and looped out bases and mini-stem loops. One stem loop from a random sequence that best simulates a

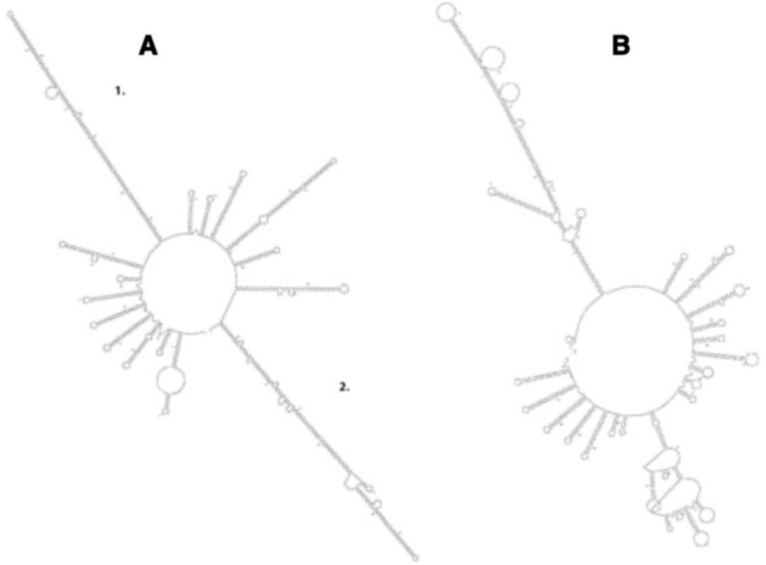


Fig. 4 Predicted DNA secondary structures generated by mfold (http://mfold.rna.albany.edu/?q=mfold/dna-folding-form). **a** Sequence from human Variable Region #3; **b** Sequence from chimpanzee Variable Region #3

Table 2 Properties of stem loops Variable Region #3

Species/Sequence	stem loop	Base pairs	TATAATATA near loop	Delta G
Human	1	109	No	−51 kcal/mol
	2	120	Yes	−52 kcal/mol
Chimpanzee	1	106	Yes	−44 kcal/mol.
Breakpoint				
Type C	1	294	Yes	−319 kcal/mol
Random				
Sequence	1[a]	123	No	−19 kcal/mol

[a]Stem loop shown in Additional file 6: Figure S5

PATRR-type structure is 123 bp in length (Additional file 6: Figure S5, stem loop 1). Thus, there is a probability that random mutations may play a role in building the secondary structures seen in the Variable Regions.

10,000 bp unit, translocation breakpoint inserts and RNA motifs

RNA motifs that reside on the 5′ flanking region of the breakpoint Type A sequence (Additional file 4: Figure S4) are found in multiple locations in the 10,000 bp chromosomal segment as a result of breakpoint insertions in the segment (Fig. 1b). These motifs consist of an exon sequence, exon 1 that is found in multiple lncRNAs transcripts (Fig. 6) and a partial sequence of an intron found in several different mRNAs primary transcripts (Fig. 7). The translocation breakpoint sequence appears to be a conduit for spreading these RNA motifs.

An analysis of base pair changes in the exon sequence alignment from lncRNAs (Fig. 6) shows that out of 212 bp, one bp change (position 101) occurs specifically in the human 10,000 bp region but none occur in the translocation breakpoint sequence, whereas other changes occur amongst the lncRNAs themselves. To a first approximation, this may indicate that the translocation breakpoint exon sequence is the original source of exon sequences found in the lncRNAs. The exon sequence found in the human 10,000 bp region shown in Fig. 6 is present in the reverse complement and is at positions 3993–4204, which is in a breakpoint sequence insert region that is not present in the comparable chimpanzee region.

Table 3 Total Variable #3 Region

Human	−257 kcal/mol
Chimpanzee	−164 kcal/mol
Breakpoint Type C	−374 kca,/mol
Random Sequence[a]	−120 kcal/mol

[a]From Fig. 4b

The mRNA intron sequence fragment (133 bp), present in the breakpoint flanking sequence is in both the human (positions 5928–6060) and chimpanzee (2801–2730) genomic regions. This sequence is found repeated in six different protein genes at ~91 % identity compared to the breakpoint sequence (Fig. 7a, b). The intron motif (here termed intron #1) is also found repeated many times in different introns of individual mRNAs, e.g., intron #1 motif is repeated 41 times in the ankyrin 2, neuronal (ANK2) gene (NCBI accession NG_009006). What stands out in the intron sequence alignment is the deletion at position 123 found only in the translocation breakpoint sequence, the human 10,000 bp sequence and the homologous chimpanzee sequence (Fig. 7a). A deletion may have occurred in the breakpoint Type A sequence before incorporation into the chimpanzee genomic region and before branching of the human species. The intron #1 sequence may have a common molecular function as it is abundantly found in several mRNAs and in multiple introns within an mRNA, however, it does not exhibit a special secondary structure or display a significant open reading frame.

Phylogenetically conserved region

Sequences of positions 5757–9694 of the human 10,000 bp segment are highly conserved between human and chimpanzee with 97 % identity over 3937 bp of the human sequence. This raises the question as to why this non-protein coding region is so well conserved. However, there are two mRNA intron sequences in this region, one of which originates from a breakpoint flanking sequence insertion. The region also displays a high nt sequence identity to an lncRNA exon (exon 4) (Fig. 1b).

One intron (intron #1), is a fragment of an mRNA intron and is at positions 5928–6060 of the 10,000 bp unit; this has been discussed above. The sequence is conserved between human and chimpanzee to ~92 % and slightly less between human and the six mRNA introns (~87 %-90 %) (Fig. 7b).

A second intron sequence (intron #2) is at positions 6670–7897 of the human 10,000 bp segment and is highly conserved between human and chimpanzee sequences (99 %) (Fig. 8). This intron is found in the RAD51 paralog B (RAD51B) mRNA. The protein is a member of the RAD51 family proteins involved in DNA repair [23]. There is an identity of 78 % between RAD51B mRNA intron and both the human and chimpanzee sequences, but the intron sequence is not found in the translocation breakpoint flanking sequence. There are multiple copies of this intron within the RAD51B mRNA itself, but the sequence is only moderately conserved within the RAD51B mRNA (75 %-80 %). The sequence is found in other mRNAs as well such as that of uracil phosphoribosyltransferase (FUR1)

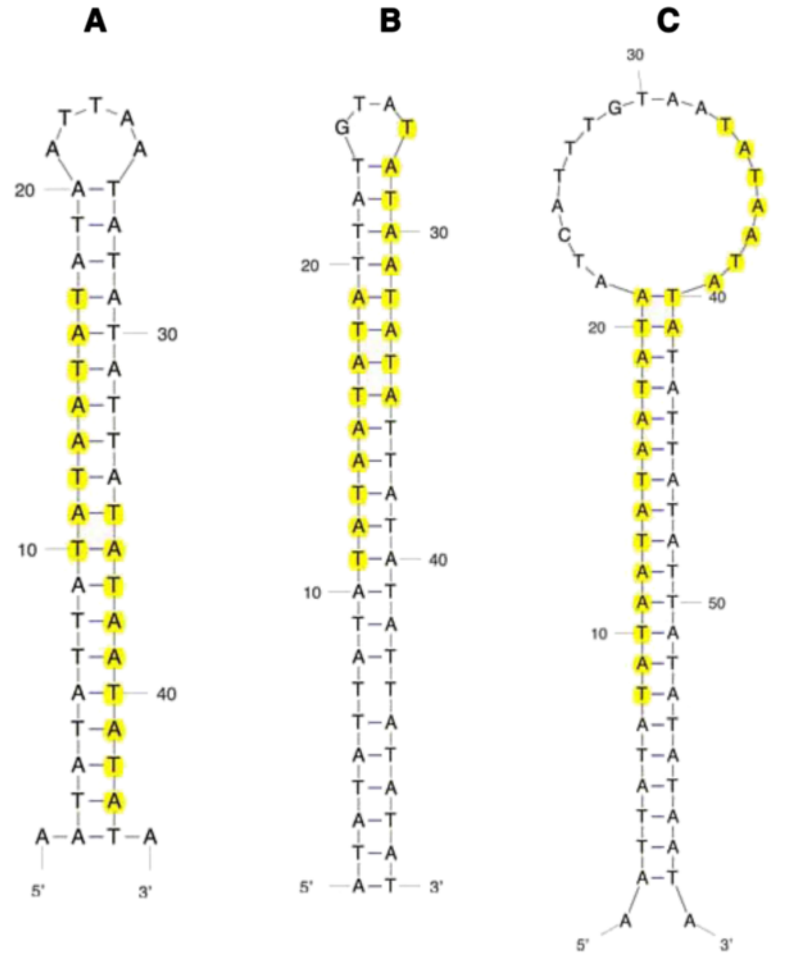


Fig. 5 Predicted apex structures of stem loops determined by mfold (http://mfold.rna.albany.edu/?q=mfold/dna-folding-form). Yellow highlighted bases signify the conserved TATAATATA motif. **a** Apex stem loop from Fig. 3a Type C PATRR, NCBI Accession number AB538237.2. **b** Apex stem loop from human stem loop 2 of Fig. 4a. **c** Apex stem loop from chimpanzee stem loop of Fig. 4b

homolog (S. cerevisiae) (UPRT) with an identity of 83 %. Thus this intron, found in several mRNAs and in the 10,000 bp region, is well conserved between humans and chimpanzee; the sequence seems to have been frozen with time between primates and humans (Fig. 8a, b).

Table 4 G:C bonds/bp

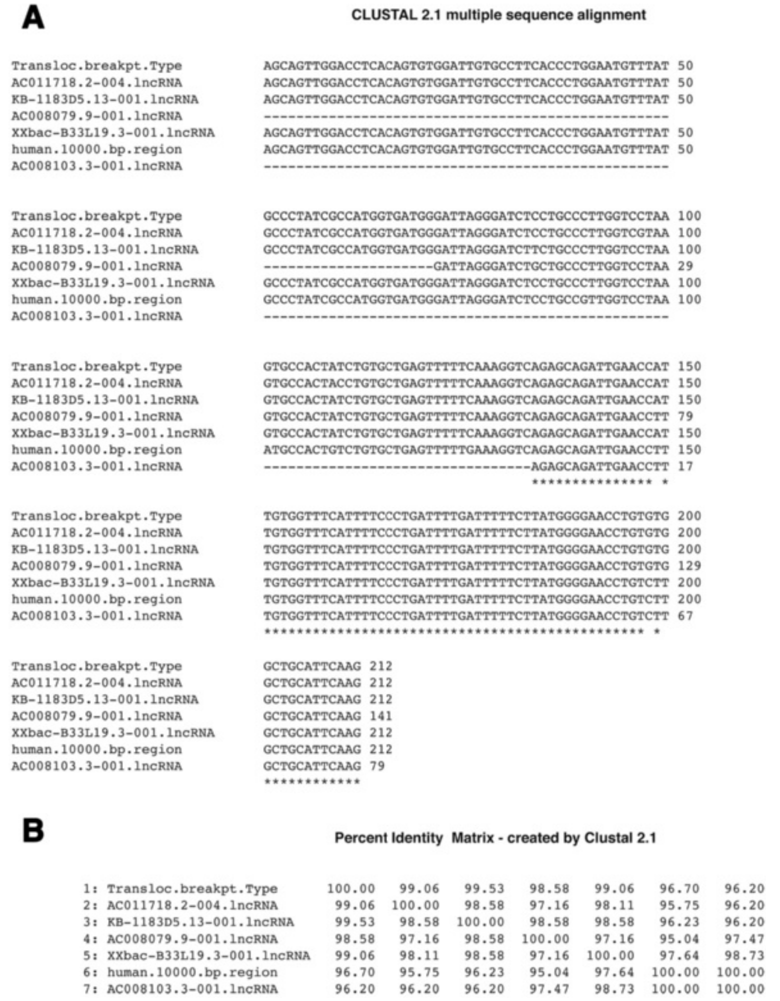
Sample	Number of G:C bonds	G:C bonds/bp
Breakpoint		
Type C	64	0.26
Human Var.#3	6	0.05
Region		
Chimp Var.#3		
Region	3	0.028
[a]Random	1	0.014

[a]From Additional file 6: Figure S5

Within the phylogenetically conserved region of the 10,000 bp unit, at human positions 8161–9421 there is a 1260 bp sequence that has an identity of 97 % with a segment of a Vega annotated lncRNA gene, AP003900.6 ENSG00000271308 (Chr. 21: 11,169,720-11,184,046). This 1260 bp sequence, found in the 10,000 bp unit includes the last exon sequence (exon 4) and short segments of flanking intron sequences of the AP003900.6 lncRNA gene. Exon 4 is also well conserved in the homologous chimpanzee region. This exon sequence adds to the variety of RNA motifs found in the 10,000 bp unit.

Positions 9423–9694 (272 bp) are well conserved between human, chimpanzee and other primates such as Rhesus Macaque but this sequence has similarities to a LINE1 element as determined by RepeatMasker [http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker].

It should be noted that Alu SINE elements have been previously found at breakpoint regions and these



A

CLUSTAL 2.1 multiple sequence alignment

```
Transloc.breakpt.Type         AGCAGTTGGACCTCACAGTGTGGATTGTGCCTTCACCCTGGAATGTTTAT 50
AC011718.2-004.lncRNA         AGCAGTTGGACCTCACAGTGTGGATTGTGCCTTCACCCTGGAATGTTTAT 50
KB-1183D5.13-001.lncRNA       AGCAGTTGGACCTCACAGTGTGGATTGTGCCTTCACCCTGGAATGTTTAT 50
AC008079.9-001.lncRNA         --------------------------------------------------
XXbac-B33L19.3-001.lncRNA     AGCAGTTGGACCTCACAGTGTGGATTGTGCCTTCACCCTGGAATGTTTAT 50
human.10000.bp.region         AGCAGTTGGACCTCACAGTGTGGATTGTGCCTTCACCCTGGAATGTTTAT 50
AC008103.3-001.lncRNA         --------------------------------------------------


Transloc.breakpt.Type         GCCCTATCGCCATGGTGATGGGATTAGGGATCTCCTGCCCTTGGTCCTAA 100
AC011718.2-004.lncRNA         GCCCTATCGCCATGGTGATGGGATTAGGGATCTCCTGCCCTTGGTCGTAA 100
KB-1183D5.13-001.lncRNA       GCCCTATCGCCATGGTGATGGGATTAGGGATCTTCTGCCCTTGGTCCTAA 100
AC008079.9-001.lncRNA         ---------------------GATTAGGGATCTGCTGCCCTTGGTCCTAA 29
XXbac-B33L19.3-001.lncRNA     GCCCTATCGCCATGGTGATGGGATTAGGGATCTCCTGCCCTTGGTCCTAA 100
human.10000.bp.region         GCCCTATCGCCATGGTGATGGGATTAGGGATCTCCTGCCGTTGGTCCTAA 100
AC008103.3-001.lncRNA         --------------------------------------------------


Transloc.breakpt.Type         GTGCCACTATCTGTGCTGAGTTTTTCAAAGGTCAGAGCAGATTGAACCAT 150
AC011718.2-004.lncRNA         GTGCCACTACCTGTGCTGAGTTTTTCAAAGGTCAGAGCAGATTGAACCAT 150
KB-1183D5.13-001.lncRNA       GTGCCACTATCTGTGCTGAGTTTTTCAAAGGTCAGAGCAGATTGAACCAT 150
AC008079.9-001.lncRNA         GTGCCACTATCTGTGCTGAGTTTTTCAAAGGTCAGAGCAGATTGAACCTT 79
XXbac-B33L19.3-001.lncRNA     GTGCCACTATCTGTGCTGAGTTTTTCAAAGGTCAGAGCAGATTGAACCAT 150
human.10000.bp.region         ATGCCACTGTCTGTGCTGAGTTTTTGAAAGGTCAGAGCAGATTGAACCTT 150
AC008103.3-001.lncRNA         ---------------------------------AGAGCAGATTGAACCTT 17
                                                               *************** *

Transloc.breakpt.Type         TGTGGTTTCATTTTCCCTGATTTTGATTTTTCTTATGGGGAACCTGTGTG 200
AC011718.2-004.lncRNA         TGTGGTTTCATTTTCCCTGATTTTGATTTTTCTTATGGGGAACCTGTGTG 200
KB-1183D5.13-001.lncRNA       TGTGGTTTCATTTTCCCTGATTTTGATTTTTCTTATGGGGAACCTGTGTG 200
AC008079.9-001.lncRNA         TGTGGTTTCATTTTCCCTGATTTTGATTTTTCTTATGGGGAACCTGTGTG 129
XXbac-B33L19.3-001.lncRNA     TGTGGTTTCATTTTCCCTGATTTTGATTTTTCTTATGGGGAACCTGTCTT 200
human.10000.bp.region         TGTGGTTTCATTTTCCCTGATTTTGATTTTTCTTATGGGGAACCTGTCTT 200
AC008103.3-001.lncRNA         TGTGGTTTCATTTTCCCTGATTTTGATTTTTCTTATGGGGAACCTGTCTT 67
                              *********************************************** *

Transloc.breakpt.Type         GCTGCATTCAAG 212
AC011718.2-004.lncRNA         GCTGCATTCAAG 212
KB-1183D5.13-001.lncRNA       GCTGCATTCAAG 212
AC008079.9-001.lncRNA         GCTGCATTCAAG 141
XXbac-B33L19.3-001.lncRNA     GCTGCATTCAAG 212
human.10000.bp.region         GCTGCATTCAAG 212
AC008103.3-001.lncRNA         GCTGCATTCAAG 79
                              ************
```

B

Percent Identity Matrix - created by Clustal 2.1

```
  1: Transloc.breakpt.Type      100.00   99.06   99.53   98.58   99.06   96.70   96.20
  2: AC011718.2-004.lncRNA       99.06  100.00   98.58   97.16   98.11   95.75   96.20
  3: KB-1183D5.13-001.lncRNA     99.53   98.58  100.00   98.58   98.58   96.23   96.20
  4: AC008079.9-001.lncRNA       98.58   97.16   98.58  100.00   97.16   95.04   97.47
  5: XXbac-B33L19.3-001.lncRNA   99.06   98.11   98.58   97.16  100.00   97.64   98.73
  6: human.10000.bp.region       96.70   95.75   96.23   95.04   97.64  100.00  100.00
  7: AC008103.3-001.lncRNA       96.20   96.20   96.20   97.47   98.73  100.00  100.00
```

Fig. 6 a Alignment of exon 1 sequences from lncRNAs, translocation breakpoint flanking sequence and the 10,000 bp unit. Nucleotide sequences of RNA exon 1 samples taken from transcript numbers shown on the Vega/Sanger website: http://vega.sanger.ac.uk/ or the Ensembl site: http://useast.ensembl.org/. The segment of the 10,000 bp unit chosen for the exon sequence alignment (positions 3993–4204) is from a previous breakpoint insert in this region. This exon sequence aligns in the reverse complement in the 10,000 bp unit. **b** Percent identities. Sequence alignment and percent identities is by Clustal W2 (www.ebi.ac.uk/Tools/msa/clustalw2/)

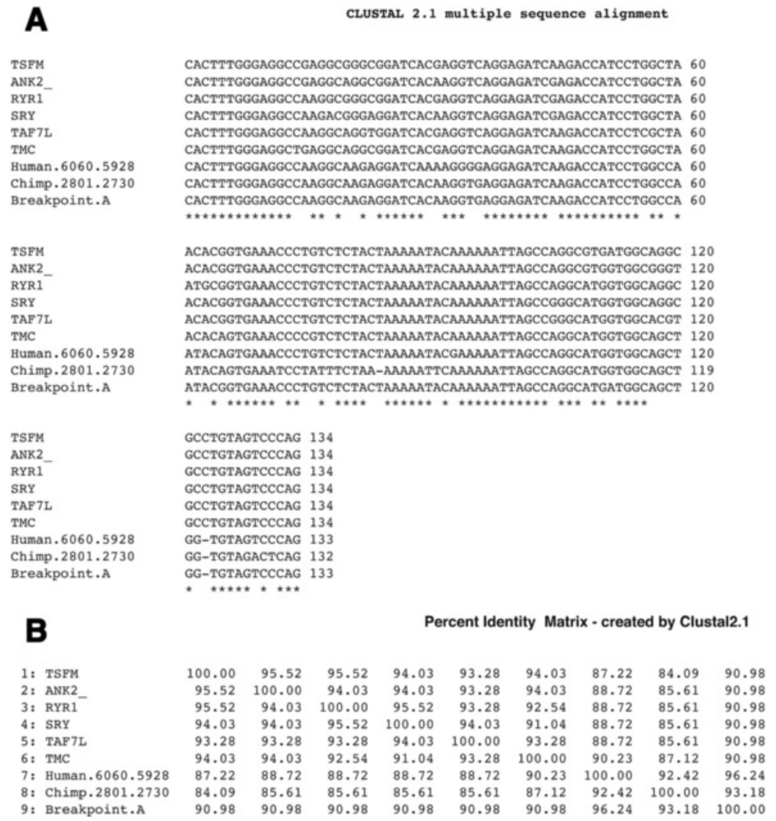
are associated with gene repeats within LCRs on 22q11.2 [24, 25]. No duplication of the highly conserved sequences (positions 5757–9694) of the human 3′ region of the 10,000 bp unit has been observed in human chromosome 22, but one copy of the 10,000 bp sequence containing the entire conserved sequence is present in each of chromosomes 13 and 21 (ND, unpublished data).

3′ end of 10,000 bp segment

A fragment of a third mRNA intron (intron #3) is present at the 3′ end of the human 10,000 bp region, positions 9700–9974 bp (274 bps), but this sequence is not present in the translocation breakpoint sequence and does not show significant identity with the chimpanzee sequence. The sequence is present in many different protein genes that includes low-density lipoprotein receptor-related protein 4 (LRP4) (Fig. 9a). There is a very high identity between the 10,000 bp unit and the seven-intron sequences shown (99-100 %) (Fig. 9b). There appears to be no obvious pattern in substitutions between the mRNA intron sequences themselves or with the 10,000 bp unit.

Translocation breakpoint sequences and linked A + T-rich regions are present in different locations of 22q11.2

Besides its presence in the 10,000 bp sequence, the unusual pattern of translocation breakpoint sequences linked to high A + T regions is also found in other regions of 22q11.2. In one example, a region was found in Chr22 (positions 18203085–18206244) that has sections of very high identity to the breakpoint type A sequence



A

```
                         CLUSTAL 2.1 multiple sequence alignment

TSFM                    CACTTTGGGAGGCCGAGGCGGGCGGATCACGAGGTCAGGAGATCAAGACCATCCTGGCTA 60
ANK2_                   CACTTTGGGAGGCCGAGGCAGGCGGATCACAAGGTCAGGAGATCGAGACCATCCTGGCTA 60
RYR1                    CACTTTGGGAGGCCAAGGCGGGCGGATCACGAGGTCAGGAGATCGAGACCATCCTGGCTA 60
SRY                     CACTTTGGGAGGCCAAGACGGGAGGATCACAAGGTCAGGAGATCGAGACCATCCTGGCTA 60
TAF7L                   CACTTTGGGAGGCCAAGGCAGGTGGATCACGAGGTCAGGAGATCAAGACCATCCTCGCTA 60
TMC                     CACTTTGGGAGGCTGAGGCAGGCGGATCACGAGGTCAGGAGATCAAGACCATCCTGGCTA 60
Human.6060.5928         CACTTTGGGAGGCCAAGGCAAGAGGATCAAAAGGGGAGGAGATCAAGACCATCCTGGCCA 60
Chimp.2801.2730         CACTTTGGGAGGCCAAGGCAAGAGGATCACAAGGTGAGGAGATCAAGACCATCCTGGCCA 60
Breakpoint.A            CACTTTGGGAGGCCAAGGCAAGAGGATCACAAGGTGAGGAGATCAAGACCATCCTGGCCA 60
                        *************  ** *  * ******  ***  ******** ********** ** *

TSFM                    ACACGGTGAAACCCTGTCTCTACTAAAAATACAAAAAATTAGCCAGGCGTGATGGCAGGC 120
ANK2_                   ACACGGTGAAACCCTGTCTCTACTAAAAATACAAAAAATTAGCCAGGCGTGGTGGCGGGT 120
RYR1                    ATGCGGTGAAACCCTGTCTCTACTAAAAATACAAAAAATTAGCCAGGCATGGTGGCAGGC 120
SRY                     ACACGGTGAAACCCTGTCTCTACTAAAAATACAAAAAATTAGCCGGGCATGGTGGCAGGC 120
TAF7L                   ACACGGTGAAACCCTGTCTCTACTAAAAATACAAAAAATTAGCCGGGCATGGTGGCACGT 120
TMC                     ACACAGTGAAACCCCGTCTCTACTAAAAATACAAAAAATTAGCCAGGCATGGTGGCAGCT 120
Human.6060.5928         ATACAGTGAAACCCTGTCTCTACTAAAAATACGAAAAATTAGCCAGGCATGGTGGCAGCT 120
Chimp.2801.2730         ATACAGTGAAATCCTATTTCTAA-AAAAATTCAAAAAATTAGCCAGGCATGGTGGCAGCT 119
Breakpoint.A            ATACGGTGAAACCCTGTCTCTACTAAAAATACAAAAAATTAGCCAGGCATGATGGCAGCT 120
                        *  * ****** **  * ****  ****** * *********** *** ** ****

TSFM                    GCCTGTAGTCCCAG 134
ANK2_                   GCCTGTAGTCCCAG 134
RYR1                    GCCTGTAGTCCCAG 134
SRY                     GCCTGTAGTCCCAG 134
TAF7L                   GCCTGTAGTCCCAG 134
TMC                     GCCTGTAGTCCCAG 134
Human.6060.5928         GG-TGTAGTCCCAG 133
Chimp.2801.2730         GG-TGTAGACTCAG 132
Breakpoint.A            GG-TGTAGTCCCAG 133
                        *  ***** * ***
```

B

```
                       Percent Identity  Matrix - created by Clustal2.1

 1: TSFM              100.00   95.52   95.52   94.03   93.28   94.03   87.22   84.09   90.98
 2: ANK2_              95.52  100.00   94.03   94.03   93.28   94.03   88.72   85.61   90.98
 3: RYR1               95.52   94.03  100.00   95.52   93.28   92.54   88.72   85.61   90.98
 4: SRY                94.03   94.03   95.52  100.00   94.03   91.04   88.72   85.61   90.98
 5: TAF7L              93.28   93.28   93.28   94.03  100.00   93.28   88.72   85.61   90.98
 6: TMC                94.03   94.03   92.54   91.04   93.28  100.00   90.23   87.12   90.98
 7: Human.6060.5928    87.22   88.72   88.72   88.72   88.72   90.23  100.00   92.42   96.24
 8: Chimp.2801.2730    84.09   85.61   85.61   85.61   85.61   87.12   92.42  100.00   93.18
 9: Breakpoint.A       90.98   90.98   90.98   90.98   90.98   90.98   96.24   93.18  100.00
```

Fig. 7 a Intron sequence alignment of protein genes, translocation breakpoint flanking sequence, the 10,000 bp unit and chimpanzee homologous region. Genomic positions of human and chimpanzee sequences are shown and are the reverse complement. Protein genes are shown in NCBI Gene Bank website: (http://www.ncbi.nlm.nih.gov/). TMC, transmembrane channel-like 1 , RefSeqGene on chromosome 9; RYR1, ryanodine receptor 1 (skeletal) (RYR1), RefSeqGene on chromosome 19; TAF7L, TAF7-like RNA polymerase II, TATA box binding protein (TBP)-associated factor, 50 kDa, RefSeqGene on chromosome X; TSFM, Ts.translation elongation factor, mitochondrial (TSFM), RefSeqGene on chromosome 12; SRY, (sex determining region Y)-box 5 (SOX5), RefSeqGene on chromosome 12; ANK2, ankyrin 2, neuronal (ANK2), RefSeqGene (LRG_327) on chromosome 4. Sequence alignment is by Clustal W2 (www.ebi.ac.uk/Tools/msa/clustalw2/). **b** Percent identities

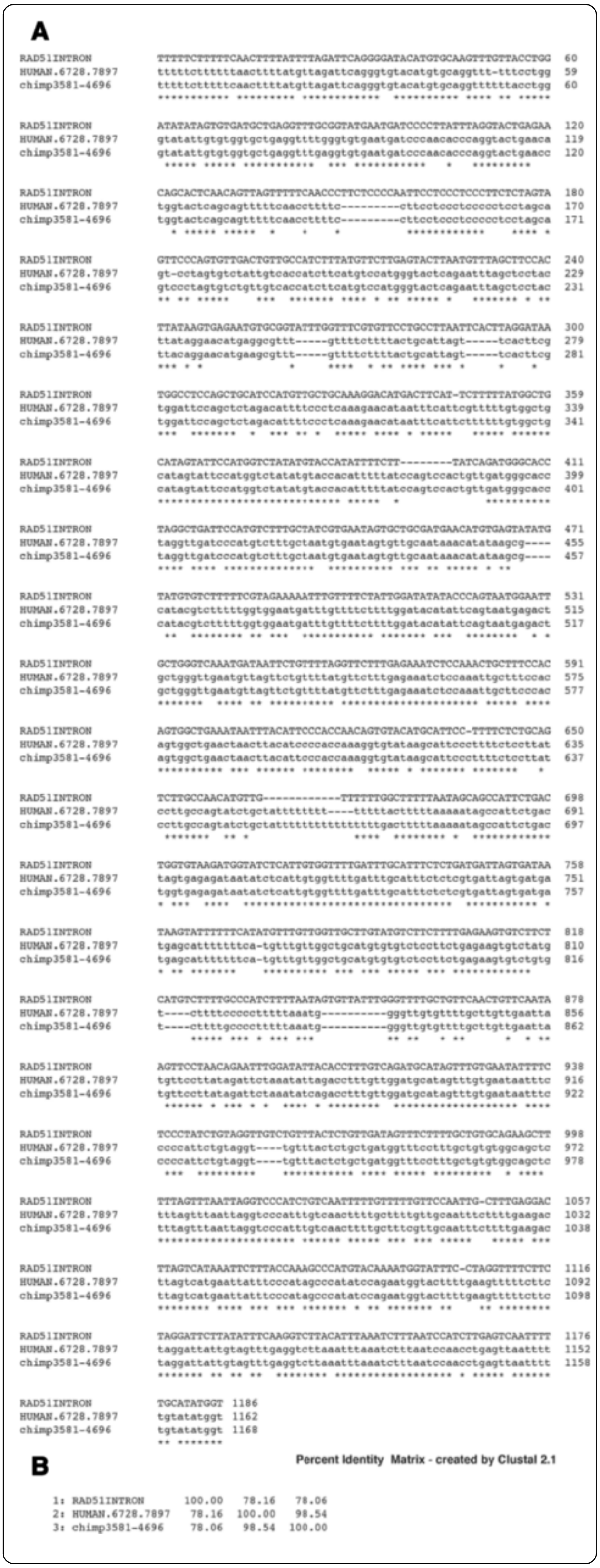
and the breakpoint sequences are linked to variable A + T-rich sequences (93 % A + T), a pattern similar to that of the 10,000 bp variable region. Figure 10 shows two breakpoint sequences (highlighted in red) that straddle a high A + T sequence (highlighted in blue).

In a similar vein to the 10,000 bp region, the high A + T-containing region has 58 copies of the sequence TATAATATA, whereas twenty-five random sample sequences of the same length and A + T content average 1.56 copies. The p-value in this case is also <0.0001.

Breakpoint sequences have been extensively duplicated in 22q11.2, and we hypothesize that they may carry information to form adjacent highly biased A + T regions that mutate extensively and are found in several regions of 22q11.2.

Discussion

Aside from protein coding regions, most of the human genome has not been defined in terms of function, although it is fascinating that much of the genome is transcribed into RNA and a large number of non-coding RNA genes have been characterized [26, 27]. In the analysis of a minute part of chromosome 22, the 10,000 bp region, there may be functions independent of regulation, expression of protein products, or RNA transcripts. With respect to RNA transcripts, no transcript with 100 % identity to annotated lncRNAs has yet been shown to originate from the 10,000 bp region (L. Wilming, Welcome Trust Sanger Institute, personal communication). However, we show here that the 10,000 bp unit has segments with highly variable sequences that appear to be reserved for biased genetic mutational processes. We hypothesize that these may lead to formation of molecular functions, e.g., PATRR formation. On the other hand, the highly conserved sequence at the 3′ half of the 10,000 bp region may be reserved for the storage of RNA motifs. The pattern, i.e., 5′ end segment reserved for mutations and 3′ end of 10,000 bp containing



A

```
RAD51INTRON         TTTTTCTTTTTCAACTTTTATTTTAGATTCAGGGGATACATGTGCAAGTTTGTTACCTGG  60
HUMAN.6728.7897     tttttctttttaacttttatgttagattcagggtgtacatgtgcaggttt-tttcctgg  59
chimp3581-4696      tttttcttttcaacttttatgttagattcagggtgtacatgtgcaggtttttttacctgg  60
                    *********** ********* ************  ********** **** ** *****

RAD51INTRON         ATATATAGTGTGATGCTGAGGTTTGCGGTATGAATGATCCCCTTATTTAGGTACTGAGAA  120
HUMAN.6728.7897     gtatattgtgtggtgctgaggttttgggtgtgaatgatcccaacacccaggtactgaaca  119
chimp3581-4696      gtatattgtgtggtgctgaggtttgaggtgtgaatgatcccaacacccaggtactgaacc  120
                     ***** ***** ***********  *** ***********   *   *********

RAD51INTRON         CAGCACTCAACAGTTAGTTTTTCAACCCTTCTCCCCAATTCCTCCCTCCCTTCTCTAGTA  180
HUMAN.6728.7897     tggtactcagcagtttttcaacctttc---------cttcctccctcccccctcctagca  170
chimp3581-4696      tggtactcagcagtttttcaacctttc---------cttcctccctcccccctcctagca  171
                      * ***** *****  *    *    *          ************    **** *

RAD51INTRON         GTTCCCAGTGTTGACTGTTGCCATCTTTATGTTCTTGAGTACTTAATGTTTAGCTTCCAC  240
HUMAN.6728.7897     gt-cctagtgtctattgtcaccatcttcatgtccatgggtactcagaatttagctcctac  229
chimp3581-4696      gtccctagtgtctgttgtcaccatcttcatgtccatgggtactcagaatttagctcctac  231
                    ** ** *****    ***  ******* **** * ** ***** *     ******* * **

RAD51INTRON         TTATAAGTGAGAATGTGCGGTATTTGGTTTCGTGTTCCTGCCTTAATTCACTTAGGATAA  300
HUMAN.6728.7897     ttataggaacatgaggcgttt-----gttttcttttactgcattagt-----tcacttcg  279
chimp3581-4696      ttacaggaacatgaagcgttt-----gttttcttttactgcattagt-----tcacttcg  281
                    *** * *             *     ****   * ** **** *** *    *    *

RAD51INTRON         TGGCCTCCAGCTGCATCCATGTTGCTGCAAAGGACATGACTTCAT-TCTTTTTATGGCTG  359
HUMAN.6728.7897     tggattccagctctagacattttccctcaaagaacataatttcattcgtttttgtggctg  339
chimp3581-4696      tggattccagctctagacattttccctcaaagaacataatttcattctttttttgtggctg  341
                    ***  *******  *  *** ** *  ***** **** * *****   ***** ******

RAD51INTRON         CATAGTATTCCATGGTCTATATGTACCATATTTTCTT--------TATCAGATGGGCACC  411
HUMAN.6728.7897     catagtattccatggtctatatgtaccacattttatccagtccactgttgatgggcacc  399
chimp3581-4696      catagtattccatggtctatatgtaccacattttatccagtccactgttgatgggcacc  401
                    **************************** *****  *             **********

RAD51INTRON         TAGGCTGATTCCATGTCTTTGCTATCGTGAATAGTGCTGCGATGAACATGTGAGTATATG  471
HUMAN.6728.7897     taggttgatcccatgtctttgctaatgtgaatagtgttgcaataaacatataagcg----  455
chimp3581-4696      taggttgatcccatgtctttgctaatgtgaatagtgttgcaataaacatataagcg----  457
                    **** **** **************  ********** *** ** ***** * **

RAD51INTRON         TATGTGTCTTTTTCGTAGAAAAATTTGTTTTCTATTGGATATATACCCAGTAATGGAATT  531
HUMAN.6728.7897     catacgtctttttggtggaatgatttgttttcttttggatacatattcagtaatgagact  515
chimp3581-4696      catacgtctttttggtggaatgatttgttttcttttggatacatattcagtaatgagact  517
                     **  ******** ** ***  *********** ******* ***  ********  * *

RAD51INTRON         GCTGGGTCAAATGATAATTCTGTTTTAGGTTCTTTGAGAAATCTCCAAACTGCTTTCCAC  591
HUMAN.6728.7897     gctgggttgaatgttagttctgttttatgttctttgagaaatctccaaattgctttccac  575
chimp3581-4696      gctgggttgaatgttagttctgttttatgttctttgagaaatctccaaattgcttcccac  577
                    *******  **** ** ********** ********************* ***** ****

RAD51INTRON         AGTGGCTGAAATAATTTACATTCCCACCAACAGTGTACATGCATTCC-TTTTCTCTGCAG  650
HUMAN.6728.7897     agtggctgaactaacttacatccccaccaaaggtgtataagcattccctttctccttat  635
chimp3581-4696      agtggctgaactaacttacattcccaccaaaggtgtataagcattccctttctccttat  637
                    ********** *** ****** ********  ***** * ******* *******   *

RAD51INTRON         TCTTGCCAACATGTTG------------TTTTTTGGCTTTTTAATAGCAGCCATTCTGAC  698
HUMAN.6728.7897     ccttgccagtatctgctatttttttt----tttttactttttaaaaatagccattctgac  691
chimp3581-4696      ccttgccagtatctgctatttttttttttttttttgactttttaaaaatagccattctgac  697
                     *******  ** *                ****  ******** *  ************

RAD51INTRON         TGGTGTAAGATGGTATCTCATTGTGGTTTTGATTTGCATTTCTCTGATGATTAGTGATAA  758
HUMAN.6728.7897     tagtgagagataatatctcattgtggttttgatttgcatttctctcgtgattagtgatga  751
chimp3581-4696      tggtgagagataatatctcattgtggttttgatttgcatttctctcgtgattagtgatga  757
                    * ***  ****  ********************************  *********** *

RAD51INTRON         TAAGTATTTTTTCATATGTTTGTTGGTTGCTTGTATGTCTTCTTTTGAGAAGTGTCTTCT  818
HUMAN.6728.7897     tgagcatttttttca-tgtttgttggctgcatgtgtgtctccttctgagaagtgtctatg  810
chimp3581-4696      tgagcatttttttca-tgtttgttggctgcatgtgtgtctccttctgagaagtgtctgtg  816
                    * ** *******    ********** *** *** ***** *** ************

RAD51INTRON         CATGTCTTTTGCCCATCTTTTAATAGTGTTATTTGGGTTTTGCTGTTCAACTGTTCAATA  878
HUMAN.6728.7897     t----cttttcccccttttaaatg----------gggttgtgttttgcttgttgaatta  856
chimp3581-4696      t----cttttgccccttttaaatg----------gggttgtgttttgcttgttgaatta  862
                         ***** *** * *** ***           ** **   * **      *  * **

RAD51INTRON         AGTTCCTAACAGAATTTGGATATTACACCTTTGTCAGATGCATAGTTTGTGAATATTTTC  938
HUMAN.6728.7897     tgttccttatagattctaaatattagacctttgttggatgcatagtttgtgaataatttc  916
chimp3581-4696      tgttccttatagattctaaatatcagacctttgttggatgcatagtttgtgaataatttc  922
                     ****** * *** * *  **** * ********  ******************* ****

RAD51INTRON         TCCCTATCTGTAGGTTGTCTGTTTACTCTGTTGATAGTTTCTTTTGCTGTGCAGAAGCTT  998
HUMAN.6728.7897     ccccattctgtaggt----tgtttactctgctgatggtttcctttgctgtgtggcagctc  972
chimp3581-4696      ccccattctgtaggt----tgtttactctgctgatggtttcctttgctgtgtggcagctc  978
                     ***  *********    *********** **** ***** *********  * ****

RAD51INTRON         TTTAGTTTAATTAGGTCCCATCTGTCAATTTTTGTTTTTGTTCCAATTG-CTTTGAGGAC  1057
HUMAN.6728.7897     tttagtttaattaggtcccatttgtcaacttttgcttttgttgcaatttcttttgaagac  1032
chimp3581-4696      tttagtttaattaggtcccatttgtcaacttttgctttcgttgcaatttcttttgaagac  1038
                    ********************* ****** ***** *** *** *****   ***** ***

RAD51INTRON         TTAGTCATAAATTCTTTACCAAAGCCCATGTACAAAATGGTATTTC-CTAGGTTTTCTTC  1116
HUMAN.6728.7897     ttagtcatgaattatttcccatagcccatatccagaatggtacttttgaagtttttcttc  1092
chimp3581-4696      ttagtcatgaattatttcccatagcccatatccagaatggtacttttgaagtttttcttc  1098
                    ******** **** *** *** ******* * ** ******* **    ** ********

RAD51INTRON         TAGGATTCTTATATTTCAAGGTCTTACATTTAAATCTTTAATCCATCTTGAGTCAATTTT  1176
HUMAN.6728.7897     taggattattgtagtttgaggtcttaaatttaaatctttaatccaacctgagttaatttt  1152
chimp3581-4696      taggattattgtagtttgaggtcttaaatttaaatctttaatccaacctgagttaatttt  1158
                    ******* ** ** **  ******** ****************** * ***** ******

RAD51INTRON         TGCATATGGT  1186
HUMAN.6728.7897     tgtatatggt  1162
chimp3581-4696      tgtatatggt  1168
                    ** *******
```

B

Percent Identity Matrix - created by Clustal 2.1

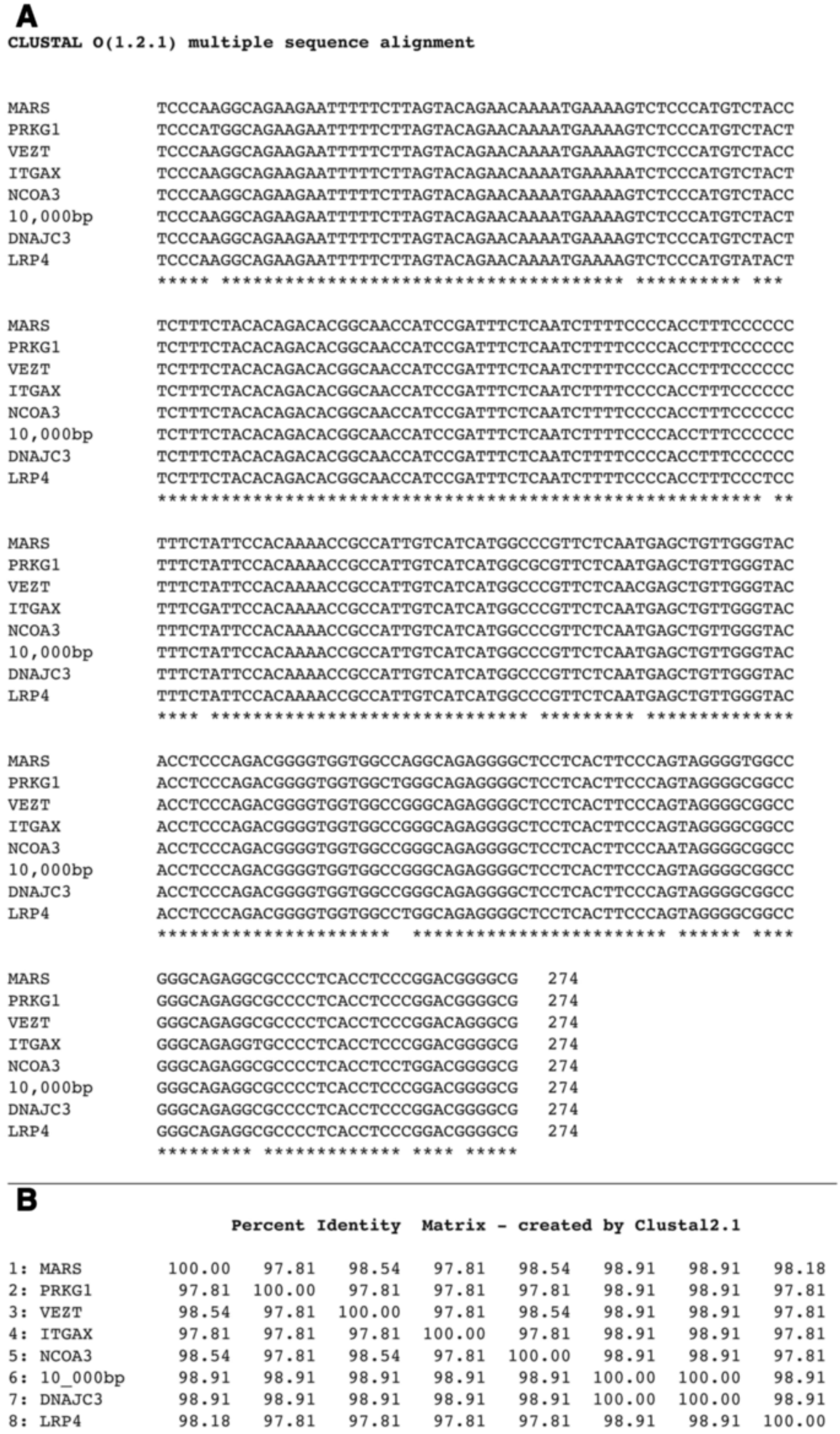
```
  1: RAD51INTRON       100.00   78.16   78.06
  2: HUMAN.6728.7897    78.16  100.00   98.54
  3: chimp3581-4696     78.06   98.54  100.00
```

Fig. 8 a Alignment via Clustal Omega of nt sequences from RAD51 intron [19], human 10,000 bp unit (positions 6728–7897) and the homologous chimpanzee region (positions 3581–4696). **b** Percent identity between the three sequences

phylogenetically conserved sequences, is conserved between chimpanzee and human genomes, which supports the idea of functional roles.

The three sequence Variable Regions of the 10,000 bp unit, consisting of approximately 1000 bp each have mutations that may be viewed as a trial and error processes with a strong partiality to maintaining a very high A + T content and formation and maintenance of many copies of the TATAATATA motif. Conceptually, this can be viewed as a process analogous to a "biased random walk" as in chemotaxis [28–30]. The analogy is that some mutations eliminate a TATAATATA unit in the human segment (e.g., see Fig. 2) but a larger number of mutations are involved in forming the unit (Table 1). As an accurate alignment of sequences within most of the Variable Regions is not possible due to the presence of low complexity sequences, we can not determine the specific number of TATAATATA units added by base pair mutations or by sequence addition, and the number lost relative to the chimpanzee sequence. Although a function for the conserved TATAATATA motif in translocation has not been determined, what is shown here is a strong bias towards forming and/or maintaining the motif with significantly more copies created in the human genome relative to that in the chimpanzee. Cairns and coworkers [31] originally proposed the concept of non-random or directed mutations. Recently, Martincorena, Luscombe and coworkers [32, 33], by using phylogenetic and population genetic methods provide evidence for non-random mutation rates. The proposal here on Variable Region mutations is consistent with these previously formed concepts, but as a biased random process. A major question remains, and that is what is the molecular mechanism that drives the mutational process towards this putative biased random mutational process? The variable regions are all linked to translocation breakpoint flanking sequences, thus, it is possible the information lies in these flanking sequences, or in the variable sequences themselves with perhaps a mechanism analogous to microsatellite replication [34].

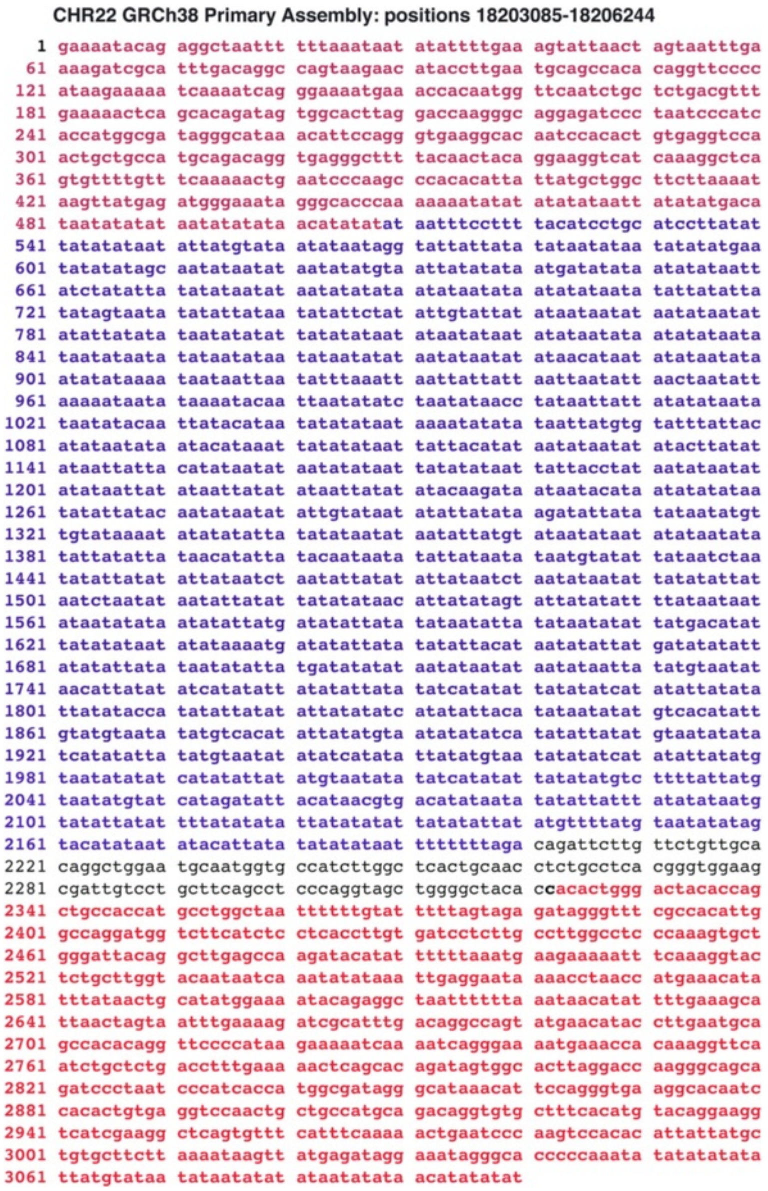
The translocation breakpoint 5′ end sequence flanking the stem loop hot spot seems to be a major player in carrying and spreading exon and intron motifs in the genome. Jacob intuitively suggested decades ago that building blocks are used over and over in biological processes to create new functions [35]. What we see here are exon and intron sequences that are used by many different genes and are found in different and multiple lncRNAs and mRNAs, respectively.



```
CLUSTAL O(1.2.1) multiple sequence alignment


MARS            TCCCAAGGCAGAAGAATTTTTCTTAGTACAGAACAAAATGAAAAGTCTCCCATGTCTACC    60
PRKG1           TCCCATGGCAGAAGAATTTTTCTTAGTACAGAACAAAATGAAAAGTCTCCCATGTCTACT    60
VEZT            TCCCAAGGCAGAAGAATTTTTCTTAGTACAGAACAAAATGAAAAGTCTCCCATGTCTACC    60
ITGAX           TCCCAAGGCAGAAGAATTTTTCTTAGTACAGAACAAAATGAAAAATCTCCCATGTCTACT    60
NCOA3           TCCCAAGGCAGAAGAATTTTTCTTAGTACAGAACAAAATGAAAAGTCTCCCATGTCTACC    60
10,000bp        TCCCAAGGCAGAAGAATTTTTCTTAGTACAGAACAAAATGAAAAGTCTCCCATGTCTACT    60
DNAJC3          TCCCAAGGCAGAAGAATTTTTCTTAGTACAGAACAAAATGAAAAGTCTCCCATGTCTACT    60
LRP4            TCCCAAGGCAGAAGAATTTTTCTTAGTACAGAACAAAATGAAAAGTCTCCCATGTATACT    60
                ***** ************************************** ********** *** 

MARS            TCTTTCTACACAGACACGGCAACCATCCGATTTCTCAATCTTTTCCCCACCTTTCCCCCC    120
PRKG1           TCTTTCTACACAGACACGGCAACCATCCGATTTCTCAATCTTTTCCCCACCTTTCCCCCC    120
VEZT            TCTTTCTACACAGACACGGCAACCATCCGATTTCTCAATCTTTTCCCCACCTTTCCCCCC    120
ITGAX           TCTTTCTACACAGACACGGCAACCATCCGATTTCTCAATCTTTTCCCCACCTTTCCCCCC    120
NCOA3           TCTTTCTACACAGACACGGCAACCATCCGATTTCTCAATCTTTTCCCCACCTTTCCCCCC    120
10,000bp        TCTTTCTACACAGACACGGCAACCATCCGATTTCTCAATCTTTTCCCCACCTTTCCCCCC    120
DNAJC3          TCTTTCTACACAGACACGGCAACCATCCGATTTCTCAATCTTTTCCCCACCTTTCCCCCC    120
LRP4            TCTTTCTACACAGACACGGCAACCATCCGATTTCTCAATCTTTTCCCCACCTTTCCCTCC    120
                ********************************************************* **

MARS            TTTCTATTCCACAAAACCGCCATTGTCATCATGGCCCGTTCTCAATGAGCTGTTGGGTAC    180
PRKG1           TTTCTATTCCACAAAACCGCCATTGTCATCATGGCGCGTTCTCAATGAGCTGTTGGGTAC    180
VEZT            TTTCTATTCCACAAAACCGCCATTGTCATCATGGCCCGTTCTCAACGAGCTGTTGGGTAC    180
ITGAX           TTTCGATTCCACAAAACCGCCATTGTCATCATGGCCCGTTCTCAATGAGCTGTTGGGTAC    180
NCOA3           TTTCTATTCCACAAAACCGCCATTGTCATCATGGCCCGTTCTCAATGAGCTGTTGGGTAC    180
10,000bp        TTTCTATTCCACAAAACCGCCATTGTCATCATGGCCCGTTCTCAATGAGCTGTTGGGTAC    180
DNAJC3          TTTCTATTCCACAAAACCGCCATTGTCATCATGGCCCGTTCTCAATGAGCTGTTGGGTAC    180
LRP4            TTTCTATTCCACAAAACCGCCATTGTCATCATGGCCCGTTCTCAATGAGCTGTTGGGTAC    180
                **** ****************************** ********* **************

MARS            ACCTCCCAGACGGGGTGGTGGCCAGGCAGAGGGGCTCCTCACTTCCCAGTAGGGGTGGCC    240
PRKG1           ACCTCCCAGACGGGGTGGTGGCTGGGCAGAGGGGCTCCTCACTTCCCAGTAGGGGCGGCC    240
VEZT            ACCTCCCAGACGGGGTGGTGGCCGGGCAGAGGGGCTCCTCACTTCCCAGTAGGGGCGGCC    240
ITGAX           ACCTCCCAGACGGGGTGGTGGCCGGGCAGAGGGGCTCCTCACTTCCCAGTAGGGGCGGCC    240
NCOA3           ACCTCCCAGACGGGGTGGTGGCCGGGCAGAGGGGCTCCTCACTTCCCAATAGGGGCGGCC    240
10,000bp        ACCTCCCAGACGGGGTGGTGGCCGGGCAGAGGGGCTCCTCACTTCCCAGTAGGGGCGGCC    240
DNAJC3          ACCTCCCAGACGGGGTGGTGGCCGGGCAGAGGGGCTCCTCACTTCCCAGTAGGGGCGGCC    240
LRP4            ACCTCCCAGACGGGGTGGTGGCCTGGCAGAGGGGCTCCTCACTTCCCAGTAGGGGCGGCC    240
                **********************  ************************ ****** ****

MARS            GGGCAGAGGCGCCCCTCACCTCCCGGACGGGGCG   274
PRKG1           GGGCAGAGGCGCCCCTCACCTCCCGGACGGGGCG   274
VEZT            GGGCAGAGGCGCCCCTCACCTCCCGGACAGGGCG   274
ITGAX           GGGCAGAGGTGCCCCTCACCTCCCGGACGGGGCG   274
NCOA3           GGGCAGAGGCGCCCCTCACCTCCTGGACGGGGCG   274
10,000bp        GGGCAGAGGCGCCCCTCACCTCCCGGACGGGGCG   274
DNAJC3          GGGCAGAGGCGCCCCTCACCTCCCGGACGGGGCG   274
LRP4            GGGCAGAGGCGCCCCTCACCTCCCGGACGGGGCG   274
                ********* ************* **** *****
```

```
                    Percent Identity  Matrix - created by Clustal2.1

1: MARS         100.00   97.81   98.54   97.81   98.54   98.91   98.91   98.18
2: PRKG1         97.81  100.00   97.81   97.81   97.81   98.91   98.91   97.81
3: VEZT          98.54   97.81  100.00   97.81   98.54   98.91   98.91   97.81
4: ITGAX         97.81   97.81   97.81  100.00   97.81   98.91   98.91   97.81
5: NCOA3         98.54   97.81   98.54   97.81  100.00   98.91   98.91   97.81
6: 10_000bp      98.91   98.91   98.91   98.91   98.91  100.00  100.00   98.91
7: DNAJC3        98.91   98.91   98.91   98.91   98.91  100.00  100.00   98.91
8: LRP4          98.18   97.81   97.81   97.81   97.81   98.91   98.91  100.00
```

Fig. 9 a Alignment using Clustal Omega of nt sequences from seven protein gene introns and the 10,000 bp unit (positions 9700–9974) (only seven of the the protein gene introns that have been detected are shown). Protein genes: PRKG1 protein kinase, cGMP-dependent, type I (PRKG1), RefSeqGene on chromosome 10; NCOA3, nuclear receptor coactivator 3 (NCOA3), RefSeqGene on chromosome 20; ITGAX, integrin, alpha X (complement component 3 receptor 4 subunit) (ITGAX), RefSeqGene on chromosome 16; MARS, methionyl-tRNA synthetase (MARS), RefSeqGene on chromosome 12; LRP4, low density lipoprotein receptor-related protein 4 (LRP4), RefSeqGene on chromosome 11; DNAJC3, DnaJ (Hsp40) homolog, subfamily C, member 3 (DNAJC3) gene; VEZT, vezatin, adherens junctions transmembrane protein (VEZT), RefSeqGene on chromosome 12;. **b** Percent identity between 10,000 bp unit and seven intron segments



CHR22 GRCh38 Primary Assembly: positions 18203085-18206244

```
   1 gaaaatacag aggctaattt tttaaataat atattttgaa agtattaact agtaatttga
  61 aaagatcgca tttgacaggc cagtaagaac ataccttgaa tgcagccaca caggttcccc
 121 ataagaaaaa tcaaaatcag ggaaaatgaa accacaatgg ttcaatctgc tctgacgttt
 181 gaaaaactca gcacagatag tggcacttag gaccaagggc aggagatccc taatcccatc
 241 accatggcga tagggcataa acattccagg gtgaaggcac aatccacact gtgaggtcca
 301 actgctgcca tgcagacagg tgagggcttt tacaactaca ggaaggtcat caaaggctca
 361 gtgttttgtt tcaaaaactg aatcccaagc ccacacatta ttatgctggc ttcttaaaat
 421 aagttatgag atgggaaata gggcacccaa aaaaatatat atatataatt atatatgaca
 481 taatatatat aatatatata acatatatat aatttccttt tacatcctgc atccttatat
 541 tatatataat attatgtata atataatagg tattattata tataatataa tatatatgaa
 601 tatatatagc aatataatat aatatatgta attatatata atgatatata atatataatt
 661 atctatatta tatataatat aatatatata atataatata atatataata tattatatta
 721 tatagtaata tatattataa tatattctat attgtattat ataataatat aatataatat
 781 atattatata taatatatat tatatataat ataatataat atataatata atatataata
 841 taatataata tataatataa tataatatat aatataatat ataacataat atataatata
 901 atatataaaa taataattaa tatttaaatt aattattatt aattaatatt aactaatatt
 961 aaaaataata taaaatacaa ttaatatatc taatataacc tataattatt atatataata
1021 taatatacaa ttatacataa tatatataat aaaatatata taattatgtg tatttattac
1081 atataatata atacataaat tatatataat tattacatat aatataatat atacttatat
1141 ataattatta catataatat aatatataat tatatataat tattacctat aatataatat
1201 atataattat ataattatat ataattatat atacaagata ataatacata atatatataa
1261 tatattatac aatataatat attgtataat atattatata agatattata tataatatgt
1321 tgtataaaat atatatatta tatataatat aatattatgt ataatataat atataatata
1381 tattatatta taacatatta tacaataata tattataata taatgtatat tataatctaa
1441 tatattatat attataatct aatattatat attataatct aatataatat tatatattat
1501 aatctaatat aatattatat tatatataac attatatagt attatatatt ttataataat
1561 ataatatata atatattatg atatattata tataatatta tataatatat tatgacatat
1621 tatatataat atataaaatg atatattata tatattacat aatatattat gatatatatt
1681 atatattata taatatatta tgatatatat aatataatat aatataatta tatgtaatat
1741 aacattatat atcatatatt atatattata tatcatatat tatatatcat atattatata
1801 ttatatacca tatattatat attatatatc atatattaca tataatatat gtcacatatt
1861 gtatgtaata tatgtcacat attatatgta atatatatca tatattatat gtaatatata
1921 tcatatatta tatgtaatat atatcatata ttatatgtaa tatatatcat atattatatg
1981 taatatatat catatattat atgtaatata tatcatatat tatatatgtc ttttattatg
2041 taatatgtat catagatatt acataacgtg acatataata tatattattt atatataatg
2101 tatattatat tttatatata ttatatatat tatatattat atgttttatg taatatatag
2161 tacatataat atacattata tatatataat ttttttttaga cagattcttg ttctgttgca
2221 caggctggaa tgcaatggtg ccatcttggc tcactgcaac ctctgcctca cgggtggaag
2281 cgattgtcct gcttcagcct cccaggtagc tggggctaca ccacactggg actacaccag
2341 ctgccaccat gcctggctaa tttttgtat ttttagtaga gatagggttt cgccacattg
2401 gccaggatgg tcttcatctc ctcaccttgt gatcctcttg ccttggcctc ccaaagtgct
2461 gggattacag gcttgagcca agatacatat tttttaaatg aagaaaaatt tcaaaggtac
2521 tctgcttggt acaataatca aatatataaa ttgaggaata aaacctaacc atgaaacata
2581 tttataactg catatggaaa atacagaggc taatttttta aataacatat tttgaaagca
2641 ttaactagta atttgaaaag atcgcatttg acaggccagt atgaacatac cttgaatgca
2701 gccacacagg ttccccataa gaaaaatcaa aatcagggaa aatgaaacca caaaggttca
2761 atctgctctg acctttgaaa aactcagcac agatagtggc acttaggacc aagggcagca
2821 gatccctaat cccatcacca tggcgatagg gcataaacat tccagggtga aggcacaatc
2881 cacactgtga ggtccaactg ctgccatgca gacaggtgtg ctttcacatg tacaggaagg
2941 tcatcgaagg ctcagtgttt catttcaaaa actgaatccc aagtccacac attattatgc
3001 tgtgcttctt aaaataagtt atgagatagg aaatagggca cccccaaata tatatatata
3061 ttatgtataa tataatatat ataatatata acatatatat
```

Fig. 10 Nucleotide sequence from Homo sapiens chromosome 22 (GRCh38 Primary Assembly, coordinates 18203085–18206244. Shown in red highlighted lettering are sequences that represent segments of the translocation breakpoint Type A sequence. The top 1–507 positions in the figure have an identity of 97 % with positions 310–814 on translocation breakpoint Type A but is present in the reverse complement. The bottom highlighted region, positions 2323–3100 in the figure have an identity of 96 % with position 310–1088 of translocation breakpoint Type A sequence, also in the reverse complement. The sequence highlighted in blue (positions 509–2200) represents the high A + T (93 %) region. The short unhighlighted sequence in black letters has not been characterized

Some intron segments are found repeated in different introns of the same mRNA, such as in ryanodine receptor 1 (skeletal) (RYR1) mRNA that has 36 repeats of the same sequence. This intron sequence is also is carried by the breakpoint sequence. A question is why do many different genes carry the same sequence? What may be special about these intron and exon sequences is that they perform a common function in different RNAs, or within the same mRNA.

Conclusions

Sequence Variable Regions at the 5′ end of the 10,000 bp non-coding segment of 22q11.2 show a biased random pattern of mutations that produces a very high A + T content, sequences that fold into long stem loops, and a high abundance of the nine nucleotide sequence TATAATATA; these Variable Sequence Regions are invariably linked to translocation breakpoint Type A 5′ flanking segments. We hypothesize that with further stem loop development these may function in DNA translocation. The 3′ half of the 10,000 bp region consists of sequences that are highly conserved between human and chimpanzee genomes, and this region contains various RNA motifs: both protein gene intron fragments and lncRNA gene exons, and some are carried by translocation breakpoint flanking sequences. As these motifs are well conserved between the two primate species and are found in multiple lncRNA or protein genes, they may be stored for future use in synthesis of new molecular functions. To our knowledge this is the first observation of translocation breakpoint sequences carrying and spreading RNA motifs and we suggest they may be reservoirs for use in formation of new lncRNA and protein genes.

In conclusion, this study defines a genomic region that is proposed to function independently of encoding protein or RNA genes, i.e., sections reserved for biased mutations and the storage of RNA motifs.

Methods

Genomic sequence searches

To find DNA nt sequences corresponding to the 10,000 bp human sequence (Homo sapiens chromosome 22, GRCh38 Primary Assembly) and chimpanzee homologous sequence (Pan troglodytes chromosome 22, Pan_troglodytes-2.1.4), the NCBI-Blast program was used (http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastHome).

The human and chimp blast pages were employed.

Alignment methods

Alignment of global DNA sequences was by one of three methods: 1. EMBL Clustal W2 (http://www.ebi.ac.uk/Tools/msa/clustalw2/) [14]; 2. ExPaSy LALIGN (http://embnet.vital-it.ch/software/LALIGN_form.html) [15]; and 3. Emboss Needle: (http://www.ebi.ac.uk/Tools/psa/emboss_needle/nucleotide.htm)

EMBOSS: the European Molecular Biology Open Software Suite [16]. The default settings were used in each case. All three programs gave a similar overall alignment between human and chimpanzee 10,000 bp sequences, however, it was not possible to obtain a consistent alignment within Variable Regions with the exception of some small segments.

Alignment of two or more sequences with reverse complement identities was by NCBI-Blast Align Sequences Nucleotide BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi?AGE_TYPE=BlastSearch&BLAST_SPEC=blast2seq&LINK_LOC=align2seq). The NCBI Blast program was used to find exon and intron homologs to those found in the translocation breakpoint flanking sequence and 10,000 bp unit.

LncRNA searches

Search for lncRNAs in chromosome 22 22q11.2 region was with three web servers: 1. Vega/Sanger website (http://vega.sanger.ac.uk/Homo_sapiens/Location/Chromosome?r=22), 2. Ensembl/Wellcome Trust/Sanger Institute/European Bioinformatics Institute: (http://useast.ensembl.org/Homo_sapiens/Location/View?db=core;g=ENSG00000197421;r=22:18782111–18812514;t=ENST00000430306), and 3. UCSC Genome Broswer: (http://genome-euro.ucsc.edu/cgi-bin/hgTracks?db=hg19&position=chr22%3A18890337-18900336&hgsid=198277878_3neHZpMdbjv5a37IhFXAN7DDTgGb).

Generated random sequences

Random sequences were generated by Molbiol.ru and adding the nucleotide length and % GC-content: (http://molbiol.ru/eng/scripts/01_16.html) Reverse complement sequences were by using the Sequence Manipulation Suite (http://www.bioinformatics.org/sms2/rev_comp.html).

DNA secondary structure modeling

Folding of DNA sequences for secondary structure features was with the mFold Web Server: (http://www.bioinformatics.org/sms2/rev_comp.html). Standard conditions (default setting) of folding temperature, ionic conditions and constraint values as were employed [36, 37]. Structures 1 which display the lowest delta G values were chosen.

Alu SINE and LINE-1 searches

RepeatMasker [http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker] was used to determine presence of Non-LTR retrotransposable elements Alu and LINE-1 in the 10,000 bp region of 22q.11.2.

p-value determinations

p-values for the determined copy number of TATAA-TATA units in a chromosomal segment vs. the copy numbers of random generated sequences was determined by a conservative nonparametric test, Wilcoxon's signed rank test [22].

Additional files

Additional file 1: Figure S1. Alignment of sequences from Homo sapiens chromosome 22 (GRCh38 Primary Assembly, coordinates:18890337–18900336) and Pan troglodytes chromosome 22, Pan_troglodytes-2.1.4, (coordinates: 17300774 to 17307562). Alignment was by Clustal W2. Breakpoint sequences are missing in the chimpanzee, although some small segments are present, e.g., the eighteen nucleotide sequence in the chimpanzee (positions 1367–1384) that are identical to the human breakpoint sequence. (PDF 30 kb)

Additional file 2: Figure S2. Secondary structure that yields high frequency of translocation. Nucleotide sequence is from PATRR Type C, accession AB 538237. Secondary structure determined by DNA mfold program [36]. (JPEG 48 kb)

Additional file 3: Figure S3. Example of a secondary structure that yields very low frequency of translocation. Sequence is from accession AF 391128 on chromosome 11. Secondary structure determined by DNA mfold program [36]. (JPEG 164 kb)

Additional file 4: Figure S4. Translocation breakpoint and flanking sequences, Type A from Accession GenBank: AB261997.1. Color code: red, intron sequence; blue, exon 1 sequence; yellow, translocation breakpoint hot spot stem loop sequence (PATRR). (JPEG 502 kb)

Additional file 5: Table S1. Folded random sequences- stem bp, lengths. (TIFF 303 kb)

Additional file 6: Figure S5. Longest and best formed stem loop (stem loop 1) from a sample of 40 random sequences. Random sequence of 1393 nt, 8.3 % G + C. Sequence folded by mfold [36]. (DOCX 53 kb)

Competing interests

The author declares that he has no competing interests.

Acknowledgements

I am grateful to Dr. Laurens Wilming of the Welcome Trust/Sanger Institute Havana manual annotation group for current information and clarifications of Vega annotated lncRNA genes, Dr. Ian Dunham of the Birney Research Group at EMBL-EBI for background information on *DGCR6L* and to Dr. Rory Johnson and Juna Carlevaro-Fita of Centre for Genomic Regulation Barcelona, Spain for searching the 10,000 bp unit for lncRNA transcripts. Dr. Beverly Emanuel, Perelman School of Medicine, University of Pennsylvania read the preliminary manuscript and provided criticism and suggestions. Dr. Jie Yang, Department of Family, Population and Preventive Medicine, Biostatistical Consulting Core, School of Medicine, Stony Brook University kindly provided *p*-values.

Funding

Partial funding provided by the Department of Molecular Genetics and Microbiology, School of Medicine, Stony Brook University, Stony Brook, NY.

Received: 26 January 2015 Accepted: 28 September 2015
Published online: 14 October 2015

References

1. Antshel KM, Kates WR, Roizen N, Fremont W, Shprintzen RJ. 22q11.2 deletion syndrome: genetics, neuroanatomy and cognitive/behavioral features keywords. Child Neuropsychol. 2005;11(1):5–19.
2. Ryan AK, Goodship JA, Wilson DI, Philip N, Levy A, Seidel H, et al. Spectrum of clinical features associated with interstitial chromosome 22q11 deletions: a European collaborative study. J Med Genet. 1997;34(10):798–804.
3. Swillen A, Vogels A, Devriendt K, Fryns JP. Chromosome 22q11 deletion syndrome: update and review of the clinical features, cognitive-behavioral spectrum, and psychiatric complications. Am J Med Genet. 2000;97(2):128–35.
4. Fine SE, Weissman A, Gerdes M, Pinto-Martin J, Zackai EH, McDonald-McGinn DM, et al. Autism spectrum disorders and symptoms in children with molecularly confirmed 22q11.2 deletion syndrome. J Autism Dev Disord. 2005;35(4):461–70.
5. Shaikh TH, Kurahashi H, Saitta SC, O'Hare AM, Hu P, Roe BA, et al. Chromosome 22-specific low copy repeats and the 22q11.2 deletion syndrome: genomic organization and deletion endpoint analysis. Hum Mol Genet. 2000;9(4):489–501.
6. Spiteri E, Babcock M, Kashork CD, Wakui K, Gogineni S, Lewis DA, et al. Frequent translocations occur between low copy repeats on chromosome 22q11.2 (LCR22s) and telomeric bands of partner chromosomes. Hum Mol Genet. 2003;12(15):1823–37.
7. Kurahashi H, Inagaki H, Yamada K, Ohye T, Taniguchi M, Emanuel BS, et al. Cruciform DNA structure underlies the etiology for palindrome-mediated human chromosomal translocations. J Biol Chem. 2004;279(34):35377–83.
8. Babcock M, Yatsenko S, Hopkins J, Brenton M, Cao Q, de Jong P, et al. Hominoid lineage specific amplification of low-copy repeats on 22q11.2 (LCR22s) associated with velo-cardio-facial/digeorge syndrome. Hum Mol Genet. 2007;16(21):2560–71.
9. Shaikh TH, O'Connor RJ, Pierpont ME, McGrath J, Hacker AM, Nimmakayalu M, et al. Low copy repeats mediate distal chromosome 22q11.2 deletions: sequence analysis predicts breakpoint mechanisms. Genome Res. 2007;17(4):482–91.
10. Kato T, Kurahashi H, Emanuel BS. Chromosomal translocations and palindromic AT-rich repeats. Curr Opin Genet Dev. 2012;22(3):221–8.
11. Inagaki H, Ohye T, Kogo H, Tsutsumi M, Kato T, Tong M, et al. Two sequential cleavage reactions on cruciform DNA structures cause palindrome-mediated chromosomal translocations. Nat Commun. 2013;4:1592. doi:10.1038/ncomms2595.
12. Kurahashi H, Inagaki H, Hosoba E, Kato T, Ohye T, Kogo H, et al. Molecular cloning of a translocation breakpoint hotspot in 22q11. Genome Res. 2007;17(4):461–9.
13. Edelmann L, Stankiewicz P, Spiteri E, Pandita RK, Shaffer L, Lupski JR, et al. Two functional copies of the *DGCR6* gene are present on human chromosome 22q11 due to a duplication of an ancestral locus. Genome Res. 2001;11(2):208–17.
14. Higgins DG, Thompson JD, Gibson TJ. Using CLUSTAL for multiple sequence alignments. Methods Enzymol. 1996;266:383–402.
15. Huang X, Miller W. A time-efficient linear-space local similarity algorithm. Adv Appl Math. 1991;12:337–57.
16. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 2000;16(6):276–7.
17. Kogo H, Inagaki H, Ohye T, Kato T, Emanuel BS, Kurahashi H. Cruciform extrusion propensity of human translocation-mediating palindromic AT-rich repeats. Nucleic Acids Res. 2007;35(4):1198–208.
18. Tong M, Kato T, Yamada K, Inagaki H, Kogo H, Ohye T, et al. Polymorphisms of the 22q11.2 breakpoint region influence the frequency of de novo constitutional t(11;22)s in sperm. Hum Mol Genet. 2010;19(13):2630–7.
19. Zhabinskaya D, Benham CJ. Competitive superhelical transitions involving cruciform extrusion. Nucleic Acids Res. 2013;41(21):9610–21.
20. Kato T, Inagaki H, Tong M, Kogo H, Ohye T, Yamada K, et al. DNA secondary structure is influenced by genetic variation and alters susceptibility to de novo translocation. Mol Cytogenet. 2011;4:18. doi:10.1186/1755-8166-4-18.
21. Gotter AL, Nimmakayalu MA, Jalali GR, Hacker AM, Vorstman J, Conforto Duffy D, et al. A palindrome-driven complex rearrangement of 22q11.2 and 8q24.1 elucidated using novel technologies. Genome Res. 2007;17(4):470–81.
22. Wilcoxon F. Individual comparisons by ranking Individual comparisons by ranking methods. Biometrics Bulletin. 1945;1(6):80–3.
23. Thomas G, Jacobs KB, Kraft P, Yeager M, Wacholder S, Cox DG, et al. A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1). Nat Genet. 2009;41(5):579–84.
24. Babcock M, Pavlicek A, Spiteri E, Kashork CD, Ioshikhes I, Shaffer LG, et al. Shuffling of genes within low-copy repeats on 22q11 (LCR22) by Alu-mediated recombination events during evolution. Genome Res. 2003;13(12):2519–32.
25. Babcock M, Yatsenko S, Stankiewicz P, Lupski JR, Morrow BE. AT-rich repeats associated with chromosome 22q11.2 rearrangement disorders shape human genome architecture on Yq12. Genome Res. 2007;17(4):451–60.
26. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, et al. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res. 2012;22(9):1775–89.
27. Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov K, et al. Defining functional DNA elements in the human genome. Proc Natl Acad Sci U S A. 2014;111(17):6131–8.
28. Adler J, Tso WW. "Decision"-making in bacteria: chemotactic response of *Escherichia coli* to conflicting stimuli. Science. 1974;184(4143):1292–4.
29. Macnab RM, Koshland Jr DE. The gradient-sensing mechanism in bacterial chemotaxis. Proc Nat Acad Sci USA. 1972;69(9):2509–12.
30. Angelani L, Di Leonardo R. Geometrically biased random walks in bacteria-driven micro-shuttles. New Journal of Physics. 2010;12:113017.
31. Cairns J, Overbaugh J, Miller S. The origin of mutants. Nature. 1988;335(6186):142–5.
32. Martincorena I, Seshasayee AS, Luscombe NM. Evidence of non-random mutation rates suggests an evolutionary risk management strategy. Nature. 2012;485(7396):95–8.
33. Martincorena I, Luscombe NM. Non-random mutation: the evolution of targeted hypermutation and hypomutation. Bioessays. 2013;35(2):123–30.
34. Richard GF, Kerrest A, Dujon B. Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. Microbiol Mol Biol Rev. 2008;72(4):686–727.

35. Jacob F. Evolution and tinkering. Science. 1977;196(4295):1161–6.
36. Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 2003;31(13):3406–15.
37. SantaLucia Jr J. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. Proc Natl Acad Sci USA. 1998;95(4):1460–5.