Junk DNA - repetitive sequences

Repetitive DNA

Eukaryote and also human DNA contains large portion of noncoding sequences. As for the coding DNA, the noncoding DNA may be unique or in more identical or similar copies. DNA sequences with high copy numbers are then called repetitive sequences. If the copies of a sequence motif lie adjacent to each other in a block, or an array, we are speaking about tandem repeats, the repetitive sequences dispersed throughout the genome as single units flanked by unique sequence are interspersed repeats.

The nature of interspersed repeats - transposable elements

Most interspersed repeats originate by a process of transposition, which is "jumping" of a DNA segment to another place of the genome. There are essentially two types of transposable DNA elements, or transposons: DNA transposons and retrotransposons. Major classes of interspersed repeats with capacity of transposition are depicted in fig. 1.

DNA transposons

DNA transposons are regarded to be inactive in the human genome due to accumulation of mutations during vertebrate phylogenesis, so we can find only their ancient remnants or "fossils". However, the active transposon derived from the human fossil elements can be engineered with the information gathered from human and other vertebrate genomes. One example is the Sleeping Beauty transposon, which is a promising component of next generation gene therapy, due to its more specific integration site (than observed e. g. for retroviruses). How a typical DNA transposon functions? The core of the transposable element codes for an enzyme transposase. This enzyme binds to the ends of the element. The ends of the transposon are formed by inverted repeats, which can therefore exchange DNA strands and stabilize the stem-loop structure necessary for transposase action. Transposase then cuts the transposon out and ligates the resulting free chromosomal DNA ends. [Nearly identical mechanism is employed during maturation of immunoglobulin (V-D-J recombination) and TCR (T-cell receptor) genes for excision of the intervening sequences. Interestingly, the enzyme that catalyse this reaction (made from two components RAG1 and RAG2) indeed probably evolved from a transposase.] The free complex transposon-transposase binds to a specific sequence motif elsewhere in the genome, transposase cleaves the host DNA and ligates the transposon into the new place. Thus, the transposon moves by a cut-and-paste mechanism and the copy number remains stable.

Retrotransposons

Retrotransposons are most important transposable elements in the human genome. First, they are much more abundant, directly forming at least 45% of the human genome (the estimations vary, but most researchers believe, that it must be even more, since ancient retrotransposons that have been inactivated, have diverged by mutation to the point where they are unidentifiable).Second, retrotransposons are still active in the human genome.

For jumping they require cellular RNA polymerases (II or III) by which they are transcribed into RNA, while the original DNA copy is maintained at the same location. The RNA copy is

reverse-transcribed into DNA, and the DNA is inserted into the genome at a new location. Thus, these elements expand in number by a duplication (copy-and-paste) mechanism. As described for the L1 retrotransposon, process of retrotransposition is prone to various mistakes, so the new copies of a retrotransposon would be largely inactivated, because of truncation or point mutation. Because most of the transposon copies are inactive, the further expansion of the retrotransposon family is governed by the few active full-length elements. However, even if all the active elements were lost later during evolution, the genome might be literarly overrun with the fossil members of the sequence family.

Retrotransposons can be further classified as autonomous and nonautonomous. Autonomous retrotransposons are coding for proteins necessary for their transposition, although they are also dependent on host RNA polymerases and DNA repair enzymes for successful jumping. Nonautonomous retrotransposons do not code for any protein and must hijack other transposon´s enzymes to be able of transposition.

LTR retrotransposons - Endogenous retroviruses

Endogenous retroviruses, also called LTR retrotransposons, resemble proviruses of true retroviruses in the composition - they contain long terminal repeats (LTRs), gag, pol, env and prt genes, but at least one of the proteins necessary for assembly of infectious viral particles is mutated or actually missing - env in particular. Thus endogenous retroviruses can move only within cells, otherwise their life cycle is similar to infectious retroviruses, e.g. HIV virus. Although endogenous retroviruses are active in many mammals, including chimpanzee, humans currently contain only fossils (mutated and incapable of transposition), which fill about 8% of the genome. Full-length endogenous retroviruses are typically 7-9 kb long, but as in case of L1 (see below), many are truncated, especially at the 5´ end. Frequently we can find only standalone LTR, as a result of retroviral insertion and subsequent intrachromosomal recombination between the LTRs or unequal recombination of the homologous chromosomes, leading to deletion of the coding part of the retrovirus (fig. 5).

Non-LTR retrotransposons

LINE

LINEs (long interspersed nuclear elements), are autonomous retrotransposons. They comprise about 21% of the human genome. The active elements belong to the most abundant LINE-1 or L-1 family, which alone comprises 17% of the genome. Of the roughly half milion of L1s in our genome, close to 10,000 are full-length and about 100 are still capable of retrotransposition. Active L1 element is about 6 kb long and contain two open reading frames, ORF1 and ORF2. 5´UTR (untranslated region) functions also as a promoter, 3´UTR contains polyA signal. Function of ORF1 is not clear, it is only known to bind to L1 mRNA, ORF2 contains reverse transcriptase and endonuclease domain and is the enzyme responsible for integration. Life cycle of L1 begins with transcription of L1 DNA by cellular RNA polymerase II and standard maturation into mRNA molecule. The L1 mRNA is transported into cytoplasm and ORF1 translated. Then the translation is reinitiated on an internal ribosomal entry site (IRES) to translate ORF2 (uncanonical and ineffective process in eukaryotes, so only portion of L1 mRNAs get their ORF2 protein). Both proteins immediately bind to the L1 mRNA. This protein-mRNA complex is transported into nucleus. ORF2 cuts chromosomal DNA at the target site (target site is not absolutely specific like it is true for restriction endonucleases, but there is some prefererence for AT rich sequences, cleavage site

are approximately TT/AAAA). The DNA cut is unequal (creating sticky ends). Free 3´OH group on one side the cleaved DNA molecule is used by reverse transcriptase of ORF2 to prime the synthesis of the first cDNA strand (target primed reverse transcription). Detailed mechanism of second cDNA strand synthesis is still subject to discussion, but the process ends by stable integration of double stranded L1 DNA on a new place in the genome. Because of staggered DNA break made by transposon endonuclease, the integrated L1 element is flanked by target site duplication 7-20 bp (fig. 2). The reverse transcriptase is often incapable to finish first strand synthesis, resulting in 5´truncation of the newly formed copy (Fig. 3A). Reverse transcriptase also lacks proofreading (3´to 5´ endonuclease) activity, often introducing mutation into the new copy. Interestingly, L1 mRNA is expressed predominantly in meiotic and postmeiotic spermatocytes, increasing thus the L1 potencial for copy expansion (copies introduced into germ line can be inherited, as opposed to somatic transposition events).

Nonautonomous retrotransposons - SINE

SINEs short interspersed nuclear elements are typically less than 500bp long and have no protein coding potential. The main SINE family in humans is formed by Alu elements (the name is derived by their discovery based on a pair of conserved AluI restriction sites). The greater than 1 million Alu elements in the human genome account for about 11% of its mass.

Alu elements share 282 bp consensus, which is related to, and was presumably derived from the SRP (signal recognition particle) RNA subunit (called 7SL RNA). SRP is a ribonucleo-protein complex, that recognizes signal peptide, binds to it and translocates the ribosome-mRNA-nascent peptide complex to endoplasmic reticulum (ER) channel, through which the nascent protein is translocated into the ER lumen or integrated into the membrane. Alus are, like 7SL RNA gene, transcribed by RNA polymerase III. Alu RNA can bind two SRP proteins (9 and 14). Presumably, Alu can thus bind to ribosomes and by its polyA tail it can bind (if the ribosome just happens to translate LINE-1 mRNA) nascent ORF2 protein, and force ORF2 protein to reverse transcribe and integrate its RNA and not the LINE-1 mRNA (fig. 4).

Function of transposable elements

From the immediate point of view, transposons have no necessary function in the cell - called, junk DNA"; or "selfish DNA", as transposons propagate on behalf of the cellular resources. On a wider scale, the motility of the retrotransposable elements can be important for genome plasticity. Occasional insertion into genes can disrupt the gene function and cause an inherited disease (Fig. 3C). LTR and LINE elements can also change gene expression, if inserted near a gene, as LTRs and LINE 5´UTR have strong promoter activity in both directions (Fig. 3F).

Because LINE-1 retrotransposon has relatively weak polyadenylation signal, it happens that the RNA polymerase II reads through it, attaching flanking DNA sequence to L1 mRNA, which is then retrotranscribed and moved into new position. So LINE-1 can be a vector for DNA shuffling. As the retrotransposed copies of L1 are often 5´truncated, the mobilized DNA can move to new position even without any sequence of the L1 vector. This might be important for shuffling of smaller DNA fragments - like for exchange of exons among genes (Fig. 3D).

L1 retrotransposition may lead even to deletions and inversions, as depicted on fig. 3E.

Very rarely, a cellular mRNA is subject to reverse transcription and transposition by an enzyme from L1 or other retrotransposons. In this case the gene is duplicated. The new copy is called processed pseudogene, as it is derived from processed mRNA lacking introns, and is usually not functional due to missing promoter (Fig. 3B). Rarely a processed pseudogene can adopt a function under selective pressure. A well known example is pyruvate dehydrogenase gene, subunit E1alpha. This gene (PDHA1) is on X chromosome in eutherian mammals. But expression of many genes residing on X chromosome is ceased during spermatogenesis, including PDHA1, although it is essential for function of all cells. This lacking function was apparently rescued by retrotransposition - there is closely related gene PDHA2 on chromosome 4 - and this gene is intronless - a typical feature of processed pseudogenes.. Highly expressed housekeeping genes have of course higher probability of retrotransposition. We thus find many processed pseudogenes for ribosomal proteins, glycolytic enzymes, beta-actin etc. Processed pseudogenes should not be mistaken for "ordinary" pseudogenes, which arose by genomic DNA duplications (e.g. pseudogenes in the hemoglobin cluster) and retain therefore the original gene structure (exons, introns, promoter, ... although with impaired function). Several genes directly derived from a retrotransposon were discovered. The latest addition is gene Peg10 (paternally expressed 10) is derived from a LTR retrotransposon of the Ty3/gypsy family (most similar retrotransposon was found in active form in fugu fish {Takifugu rubripes}). Peg10 is necessary for placental development in mice, the same would be probably true for humans. Other examples include syncytin genes derived from endogenous retroviruses of HERV-W family. These are important in syncytia formation from trophoblast cells, mechanism of membrane fusion indeed resembles retroviral entry to cell.

Even inactive repeat elements increase plasticity of the genome by promoting interchromosomal unequal crossing-over or intrachromosomal recombination, leading to deletions/duplications or inversions (fig. 5).

Last but not least, transposons are speculated to have some real physiological function, since e.g. their expression is upregulated during stress response. But the diverse hypotheses that can be drawn from this observation are far from being elucidated.

Tandem repeats

Tandem repeats are made of succesive identical or nearly identical (degenerate) repeat units. They vary in length of repeat unit as well as lenght of the whole repeat much, so every classification is not satisfying and must be taken "cum grano salis". The largest repeats, which tend to be composed from large repeat units are called **satellites**. The name satellites comes from centrifugation of DNA in density gradients. First, during DNA isolation conventional methods, DNA is subject to shear stress, with resulting DNA fragmentation (note that in vivo one G1 phase chromosome contains 1 DNA molecule). These fragments can be then centrifuged in density gradients so the DNA molecules occupy places in the gradient with the same density as the DNA molecule. Bulk of DNA will form one band. But DNA fragments with significantly different CG/AT content, caused e. g. by large monotonous repeats will form minor "satellite" bands. The denomination of satellite DNA was later broadened to incorporate similarly repetitive sequences that are not forming these satellite bands. Satellite primary repeat units are various, from GGAAT found in satellites 2 and 3 to 171 bp in alpha satellite. But these primary units are often degenerated, conaining certain irregularities. These irregularities can be periodical, forming thus secondary repeat units. Satellite DNA is abundand at centromeres and constitutive heterochromatin. Although human genome is considered completely assembled, the centromere regions and heterochromatin containing

satellite sequences are not included, since the sequencing of such regions is from various reasons challenging (absence of restriction sites, difficult sequencing, almost impossible contig assembly). From the various satellites found at or near the centromere, a family of alpha-satellite repeat (with primary unit 171 bp) probably form functional core of centromeres, as they are important for kinetochore assembly during cell division (some kinetochore proteins bind to the alpha-satellite at centromere, and thus nucleate kinetochore assembly). The function of other satellites is unknown, regarded mostly as junk DNA.

Minisatellites are shorter tandem repeats, in the range of kb, which are enriched in subtelomeric regions of chromosomes. They are often highly polymorphic as to the number of repeat units in a repeat (many alleles in the population) and can be used as genetic markers - VNTR, variable number of tandem repeats. VNTRs are often too large to be amplified by PCR and are therefore typically assayed by Southern blot. Sometimes, certain minisatellites are hypothesised to have regulatory functions, as e.g. a VNTR in insulin promoter, where different length of the repeat was associated with different types of diabetes. One allele of the insulin VNTR is shown on fig. 7. Telomeres of human chromosomes, formed by several kilobases of the hexamer repeat TTAGGG belong also to the minisatellite range of tandem repeats, although they arise by a specific mechanism - by the enzyme telomerase. Telomerase is composed from a protein subunit with reverse transcriptase activity and an RNA subunit with a sequence complementary to TTAGGG, which serves as a template for the telomere elongation (telomerase protein subunit is related to reverse transcriptase of non-LTR retrotransposons). However, telomeres can elongate even by the passive general mechanism of unequal crossing-over (see fig. 5D), e.g. in cancer cells.

Maybe it should be noted here once again, that the sequence of the human genome comprises the euchromatic regions, bounded proximally, but not including the centromeres and pericentromeric heterochromatin, and distally by telomeres, which are also, together with subtelomeric regions not included.

Microsatellites have repeat units typically 1-5 bp, with repeat length rarely exceeding hundreds of repetitions in order. Most common family of these repeats are 2 bp repeats, from which (CA)n repeats are prevailing. The microsatellites are very common in the genome, highly polymorphic and are very often used as genetic markers. Examples of such genetic markers are in chapter covering linkage.

Trinucleotide expansion diseases

If in or near the genes, length of microsatellites can have deep consequences - e.g. in so called trinucleotide expansion diseases, a group of heterogeneous hereditary mendelian syndromes. The most known example is Huntington chorea, fatal neurological illness with adult onset presenting as dementia and extrapyramidal motion control impairment. In the huntingtin gene, there is a CAG repeat sequence, coding for a stretch of glutamine residues (polyglutamine tract) in the huntingtin protein. Normally people have less than 20 CAG trinucleotides and consequently glutamines in huntingtin, where it serves as an important domain for protein-protein interaction. However, if by mutation this number expands to more than 30 glutamines, the protein does not function properly, resulting in progressive death of neurons in nucleus caudatus. In other trinucleotide expansion disease, myotonic dystrophy (muscle dystrophy with muscle weakness paradoxically accompanied with increased muscle tone), pathologic expansion of the trinucleotide CTG takes place in the 3´ untranslated region of the DMPK (dystrophia myotonica protein kinase). The mutant mRNA itself has therefore the

pathological potential, and probably wreaks havoc through sequestration of various transcription factors. For other examples of "expansion" diseases refer to chapter Nonmendelian Inheritance.

Mechanisms of tandem repeat expansion/shrinkage

First mechanism that contributes to polymorphism of tandem repeat length is unequal crossing-over. That is typical in particular for the larger repeats (Fig. 5D). Small microsatellite repeats often change their length by mistakes of DNA synthesis, e. g. a mechanism referred as polymerase slippage (Fig. 8). At the front of replication, the DNA double helix is not yet extremely stable and is subject to substantial thermal fluctuations. If the polymerase just happens to replicate at the microsatellite, the DNA strands might not (during fluctuations) reassociate exactly, but with a shift of several repeat units. This mechanism is enhanced in some kinds of repeats that can stabilize transition states by forming double strand loops, e.g. the CAG/CTG trinucleotide.

Links

Repetitive sequences are stored in a cetral database, Repbase (unfortunately, direct use of RepBase is possible only for academic institutions). http://www.girinst.org/

There are also specialized databases, covering only some aspects, like database of human endogenous retroviruses. http://herv.img.cas.cz/

RepeatMasker is a computer program performing identification of repetitive sequences using Repbase and eventually their masking in the sequence (e.g. to facilitate gene discovery). http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker

SRPDB (signal recognition particle database) provides sequences and structures related to functions of SRP. http://psyche.uthct.edu/SRPDB/SRPDB.html

AluGene is database of Alu elements incorporated within protein-coding genes http://alugene.tau.ac.il/

L1Xplorer is a database dedicated to detection and annotation of full-length intact L1 elements http://l1xplorer.molgen.mpg.de

Links

Fig. 1: Various classes of transposons occurring in the human genome
A: Non-LTR retrotransposons. LINEs (long interspersed repeats) are represented by LINE-1 (L1). 6 kb element contains two open reading frames. The ORF2 contains endonuclease (en), reverse transcriptase (rvt) domain as well as a cysteine-rich domain (C-rich). 5´ untranslated region (5´UTR) contains also internal promoter for RNA polymerase II (in a usual gene, promoter is upstream 5´UTR). 3´ untranslated region (3´UTR) contains canonical polyadenylation signal (AATAAA) and a polyA tail (that is also normally absent from the ordinary genes, and is only added to mRNA by action of polyA polymerase). L1 is flanked by target site duplication (TSD) that arises during the target primed reverse transcription.
B: LTR-retrotransposon - endogenous retrovirus. Depicted is a typical structure of a

retrovirus, or more precisely of a provirus, the form integratred into DNA. Endogenous retroviruses can be distinguished from the infectious ones only by the means of point mutations or deletions in the genes necessary for infectious particle assembly - in most cases it is the env (envelope) gene. gag (group specific antigen) is the nucleocapsid protein. pol (polymerase) has the reverse transcriptase (rvt) activity for first and second strand DNA synthesis, RnaseH activity for cleavage of RNA in the RNA/DNA hybrid after first strand synthesis and integrase (int) activity (cleaves the target DNA and ligates the retrovirus into the cleaved site). prt (protease) is indispensable for virus assembly by cleavage protein precursors translated from retrovirus mRNA (e.g. gag and pol are often translated as one large polyprotein). LTRs (long terminal repeats) are identical sequences at the retrovirus ends. Each LTR is composed from U3 (3´ untranslated region), R (recombination region) and U5 (5´ untranslated region). This is derived from the retrovirus mRNA structure, which extends only from upstream R to downstream R. How the full length cDNA is derived from this mRNA is beyond the scope of the chapter. Although the endogenous retroviruses are reverse transcribed in cytoplasm, so the mechanisms of integration in theory does not require target site duplications, these are often formed, albeit shorter than in L1.
C: DNA transposon is represented by the 1,2 kb mariner family. The synthetic DNA transposon Sleeping Beauty belongs to this family too. Central transposase region is flanked by inverted repeats. Upon integration target site duplication is formed from the host DNA. The target site duplication is leaved in the genome as a transposon signature, when the transposon jumpes to another place.
D: nonautonomous nonLTR retrotransposons belong to the SINE (short interspersed repeat). The subfamily active in humans is represented by a typical 282 bp Alu element. Alu is a dimmer composed from two nearly identical monomers (light and intermediate grey). The left monomer has a deletion of the dark grey box. The monomer is derived from 7SL RNA gene, coding for the RNA subunit of SRP (signal recognition particle). SRP is a complex recognising signal peptide of the Sproteins that are to be transported into endoplasmic reticulum lumen or membrane. Note that the 7SL gene is drawn in 50% scale! PolyA region of Alu is not part of the 7SL gene, but is important for success of Alu in retrotransposition.

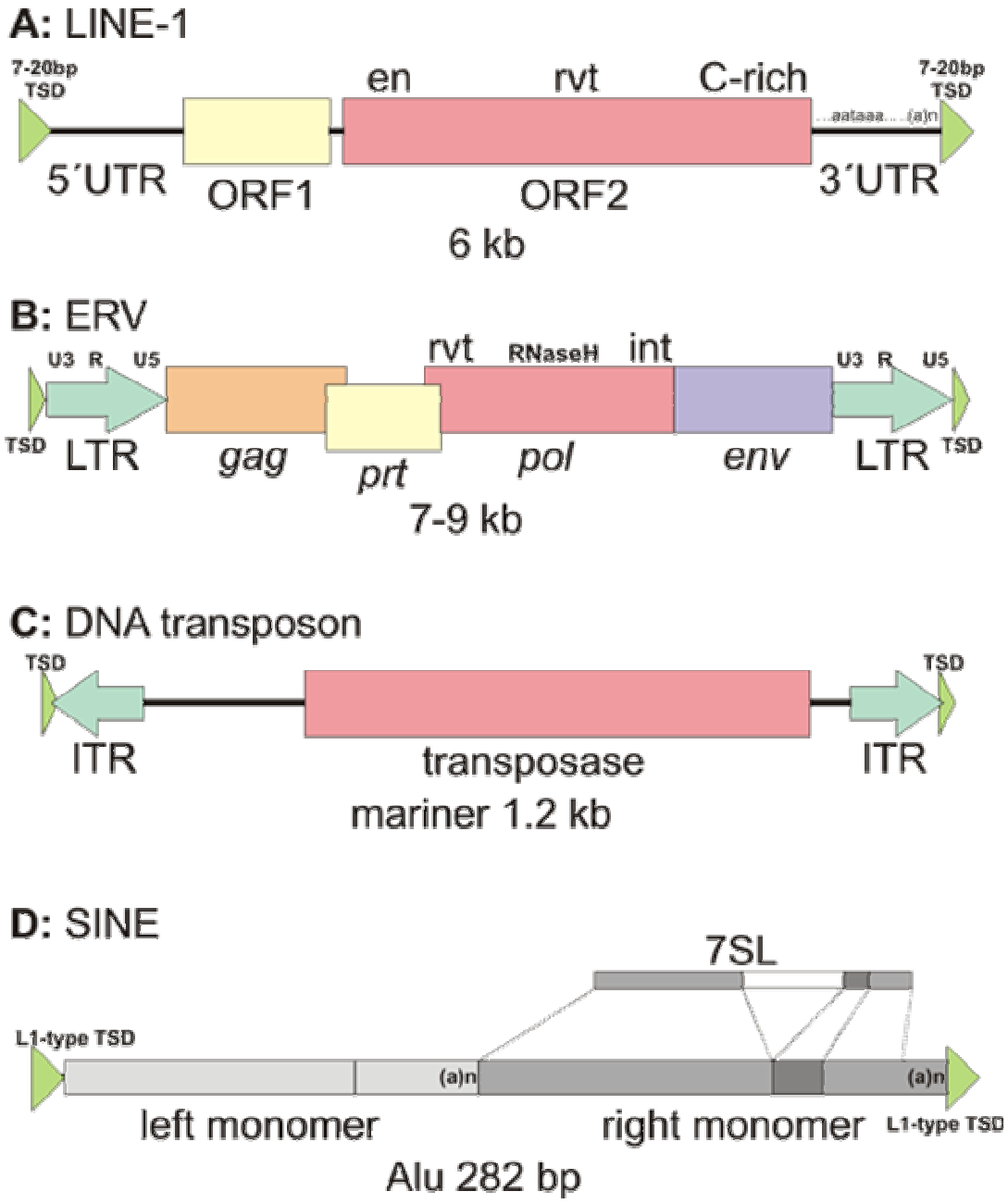


Figure 2. Target primed reverse transcription (TPRT)
ORF2 protein cleaves first one DNA strand at the target (target sequence is rich in A+T and the sequence is usually similar to consensus TTAAAA, cleavage occurs between T and A on the complementary strand). The cleaved strand dissociates and binds to polyA tail of L1 mRNA (dashed orange line). Free 3´ OH group of the DNA strand primes cDNA first strand synthesis. Cleavage of the second DNA strand occurs 7-20 nt downstream of the first cut and the free 3´ OH group generated by this event is used to prime the second strand synthesis of L1 cDNA. The mechanism of second strand synthesis is not completely elucidated. The whole process ends by formation of a new DNA copy of L1, flanked by duplication of the target site.

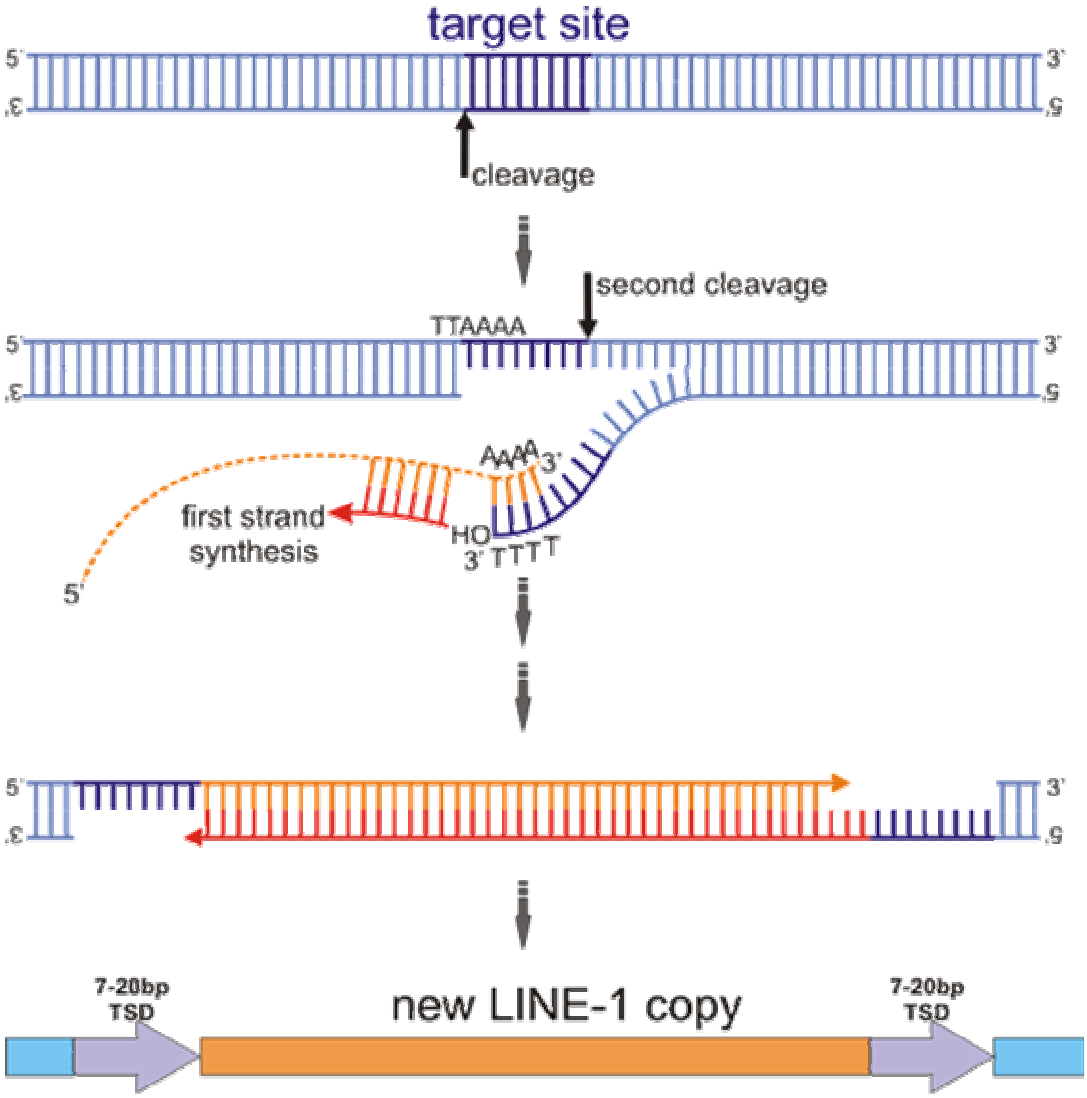


Figure 3. LINE-1 alters genome in various ways
A: Retrotransposition in cis. L1 makes retrotransposed copies of itself. The copies may be complete, or more often 5´ truncated or 5´ truncated with inversion. **B:** ORF2 protein of L1 can retrotranspose SINE elements (like Alu) or other cellular mRNAs, creating processed pseudogenes (Retrotransposition in trans). Coding exons are represented by brown boxes, 5´ and 3´ UTR (untranslated regions) are in lighter color, splicing of exons into mRNA indicated by broken lines. **C:** Retrotransposon can insert into gene. Insertion into exon usually leads to disruption of open reading frame and protein truncation (asterisk depicts the place of a new stop codon). But even insertion into an intron can have deleterious consequences - e.g. exon skipping or creation of a new exon, which also often disrupt the protein. Retrotransposon insertion is a well documented cause of various heritable diseases. Most frequently inserted are Alu elements, followed by L1. **D:** 3´ transduction. L1 has relatively weak polyadenylation signal. Therefore RNA polymerase may read through and transcribe also a segment of flanking chromosome DNA. This hybrid mRNA is then retrotransposed, resulting in moving of both L1 (which is however usually partially 5´ truncated or even completely deleted) and the flanking DNA. This may be a mechanism of exon shuffling between genes. **E:** Insertion of a retrotransposon is often accompanied by a rearrangement - here deletion of the green segment and inversion of the red segment including an exon, with subsequent skipping of this

exon during splicing. **F:** L1 promoter can promote transcription not only of its own element, but also of the neighboring genes, both upstream and downstream.

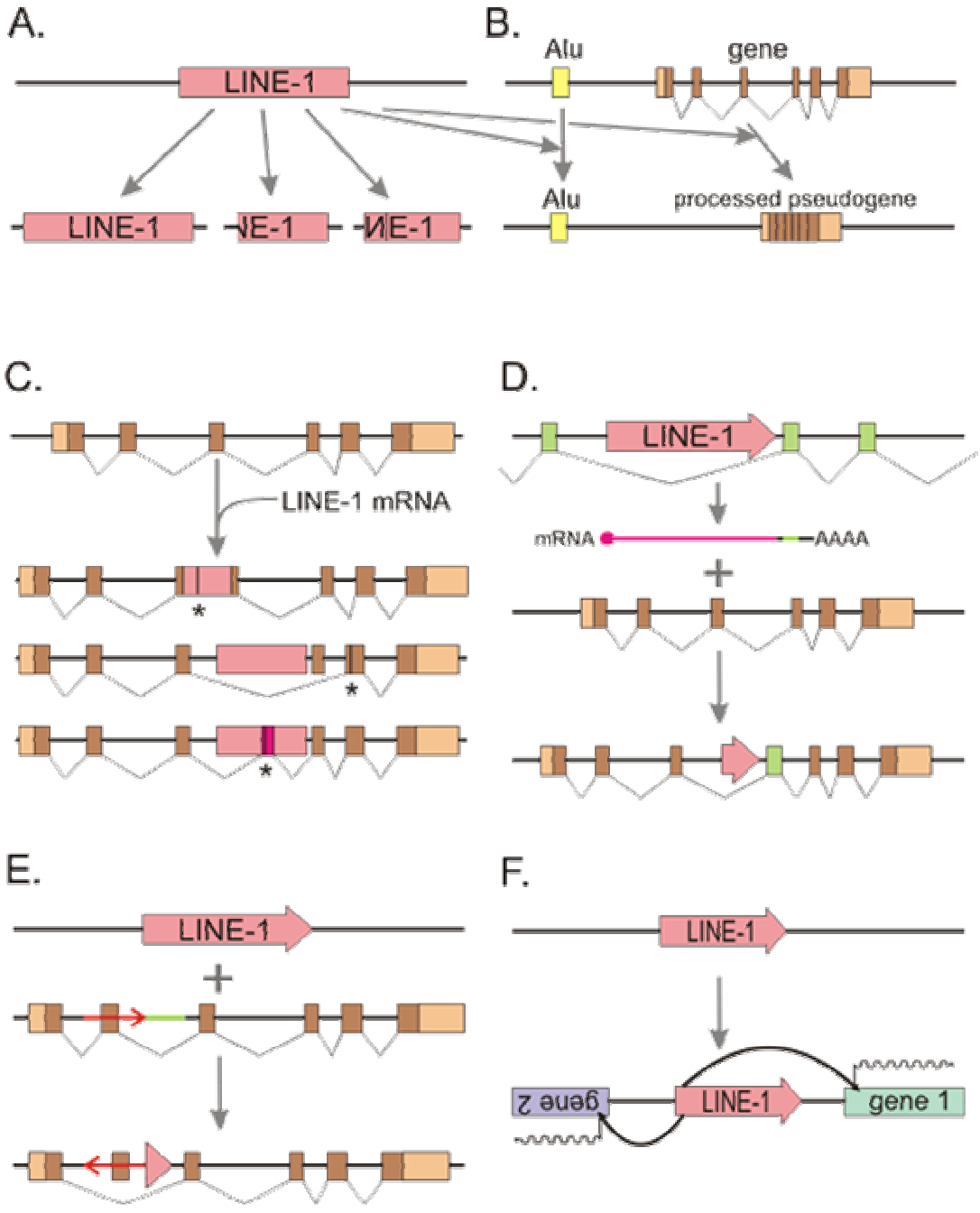


Figure 4 Alu sequences are hyperparasites
A: Structure of the 7SL RNA gene and Alu element (left) and secondary structure of the respective RNA molecules (right). Transcription of 7SL RNA gene is directed by internal RNA polymerase III promoter (A) and enhancer (EN). Alu gene has composite internal promoter (A+B). Natural terminator of RNA polymerase III is tetranucleotide TTTT. Transcription is interrupted after first three T. 7SL RNA is composed from Alu domain (blue) and S-domain (yellow). SRP proteins 9 and 14 bind to the Alu domain, which serves for the anchorage to ribosome. Other proteins bind to the S-domain, includin protein 54, which collaborates on the signal peptide (red line) recognition. Alu RNA is formed basically by two

Alu domains of 7SL RNA, with an addition of a polyA sequence.
B: Alu RNA binds to ribosome. If the ribosome is just translating ORF2 of LINE-1 mRNA (green line), the polyA tail of Alu element competes with the polyA tail of L1 for binding of nascent ORF2. PolyA binding proteins mediate the interaction. If ORF2 binds to Alu, ORF2 will reversely translate and transpose Alu instead of L1 and thus parasite on L1. If we consider L1 as a genomic parasite, Alu is a hyperparasite - i. e. parasite´s parasite. Other cellular mRNAs (blue line) can compete with the L1 mRNA for ORF2 binding too, albeit with much lower efficiency (it is estimated, that from 3000 L1 retrotranspositions, 300 would be hijacked by Alu elements and only cca 1 by another mRNA.

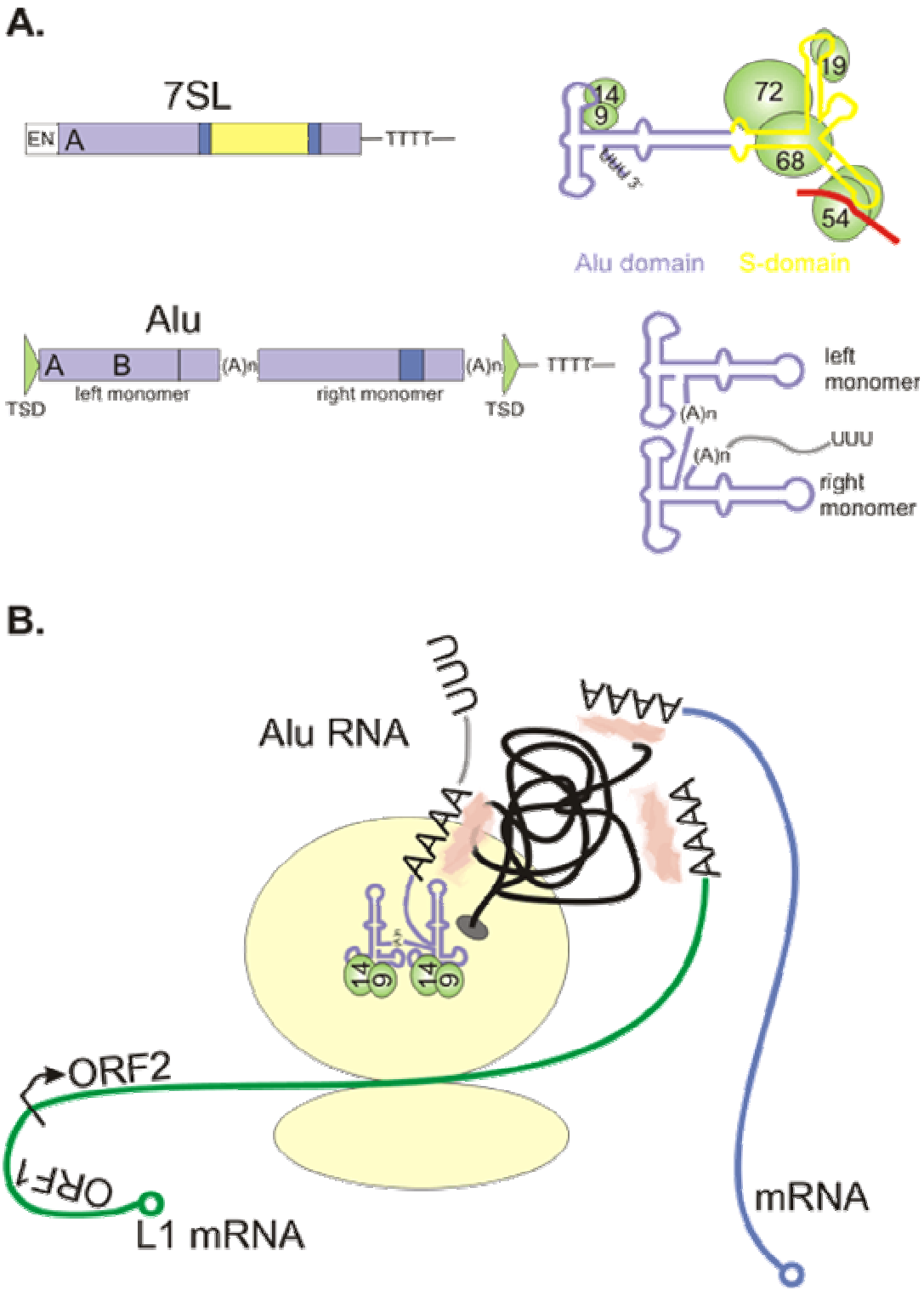


Figure 5. Repeats promote genomic rearrangements.
A+B: genomic region containing direct repeats (in the same direction, the same sequence on the same DNA strand). The two repeats can pair and recombine. The intrachromosomal recombination (A) leads to deletion. Hypothetical circular fragment is lost - it does not possess a centromere. Unequal crossing-over with resulting interchromosomal recombination (B) causes deletion and duplication.

C: Intrachromosomal recombination between two inverted repeats (in the opposite direction, the same sequence is on the opposite DNA strand) leads to inversion of the intervening DNA sequence. The functional consequences of such rearrangements are context dependent, from silent to lethal, as may be expected.
D: Tandem repeat polymorphisms can arise by unequal crossing over.

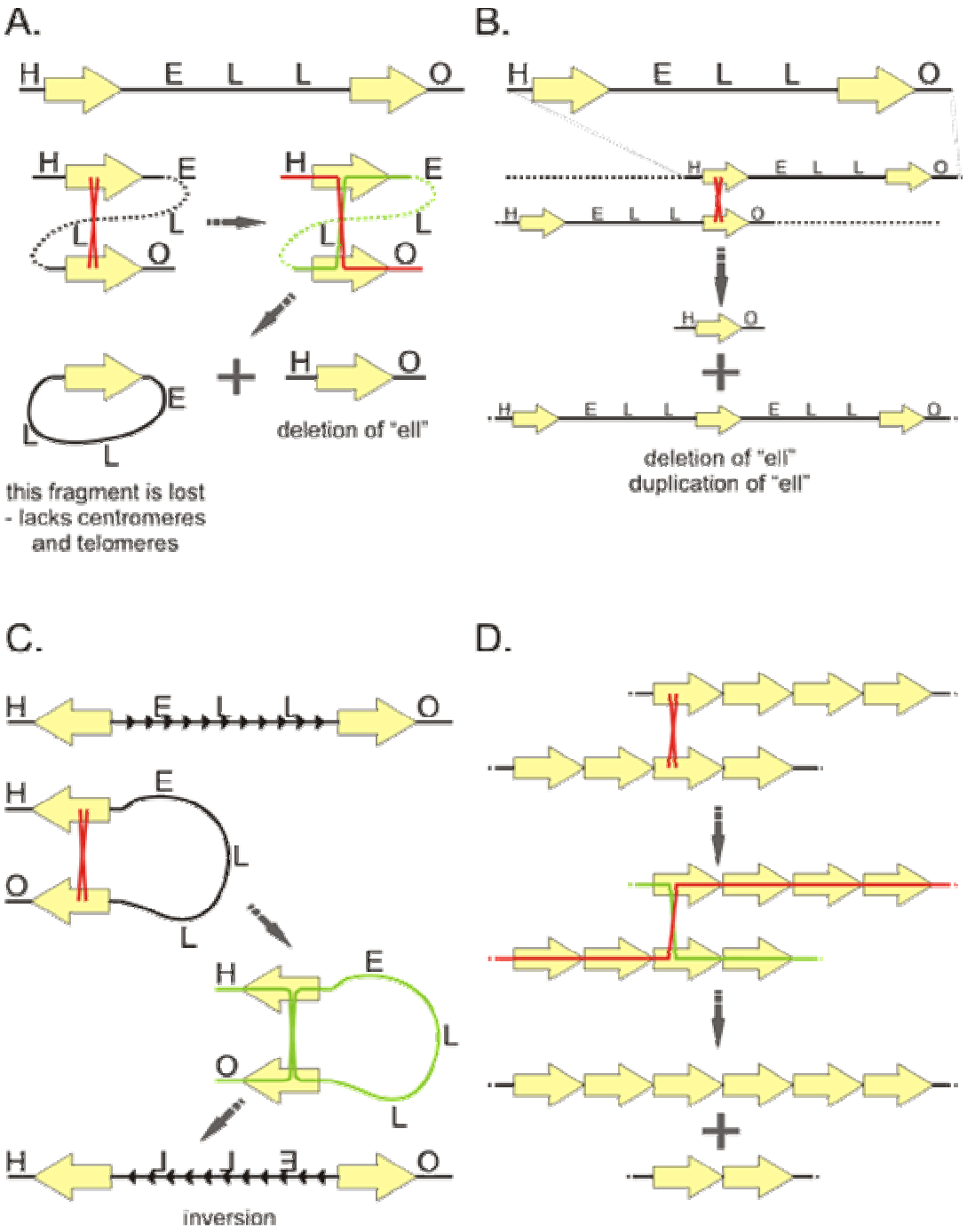
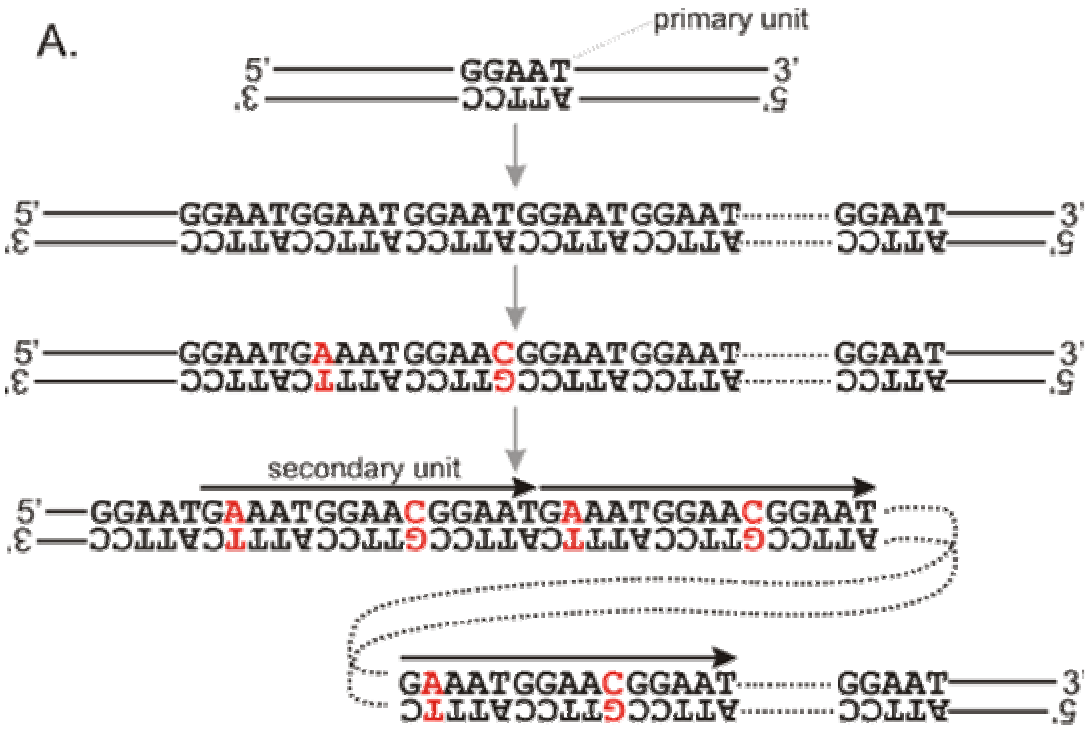


Figure 6. Satellites
A: primary units and higher order (secondary) units of tandem repeat. Probable "evolutionary history" of repeats as exemplified by GGAAT sequence. This sequence multiplicates and form thus a perfect monotone repeat. Some positions later undergo mutation (red) creating imperfect (degenerate) repeat. Then the sequence multiplicates again, but now several

degenerated units multiplicate together as one unit, creating thus a perfect repetition of this larger, secondary unit (arrow). The sequence GGAAT is base of the human satellites 2 and 3. These satellites differ by the secondary unit.
B: Structure of human mitotic chromosome with respect to satellite sequences. Alpha-satellite forms heterochromatin at the core of the centromere. Besides the proteins associated with heterochromatin, alpha-satellite binding proteins assemble on the alpha satellite sequences to form inner plate of the kinetochore. Some of these proteins are associated with the centromere throughout the cell cycle. On the inner kinetochore plate assembles an outer kinetochore plate that interacts with microtubules of the mitotic spindle. Centromere is usually flanked by pericentric heterochromatin formed by other types of satellite sequences. Tips of the chromosome (telomeres) are formed by telomeric repeat TTAGGG, the adjacent subtelomeric regions are also higly repetitive.



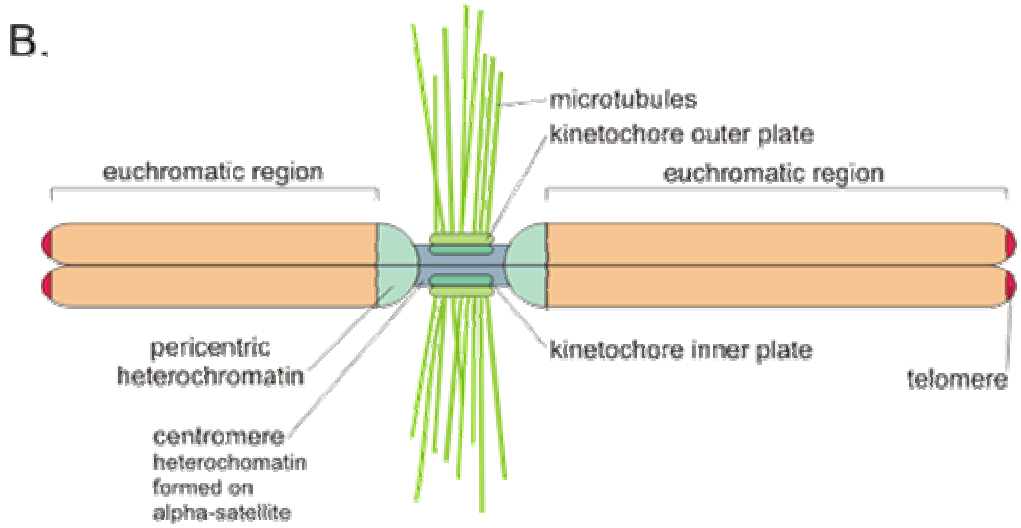


Figure 7. VNTR in insulin gene

A: DNA segment (coding strand, 5´ to 3´ direction) containing insulin gene. Insulin gene contains three exons (upper case) that form the mature mRNA. The important regulatory sequence motifs are in red - TATA box upstream of the transcription initiation site, ATG as start of translation (transcribed into AUG in mRNA, which serves as the initiation codon, inserting the first methionine of the polypeptide strand), the conserved dinucleotides GT and AG at intron splice sites, stop codon TAG and polyadenylation signal AATAAA. Sites of single nucleotide polymorphisms are in bold (that means that many subjects have a different nucleotide at that position, not the one shown). The minisatelite is in blue, of course, only one allele is shown, other alleles differ by the number of repetitions. **B:** This allele of the VNTR consists of 29 repetitions of the sequence motif GGGGTGTGGGGACA, although not all repeat units match the consensus perfectly (non-matching bases are in black). Note that the repetition contains a palindrome TGTnnnnACA, which may stabilize "stem-loop" structures and promote thus instability of the number of repeats (see fig. 8). Variable length of the minisatelite just upstream of the insulin gene in promoter region may differentially interact with transcription factor binding promoter and cause thus differential expression of the insulin gene. Indeed some alleles were associated with development of diabetes (however, is is very challenging to differentiate direct effect from "only" linkage - see the chapter dealing with linkage.

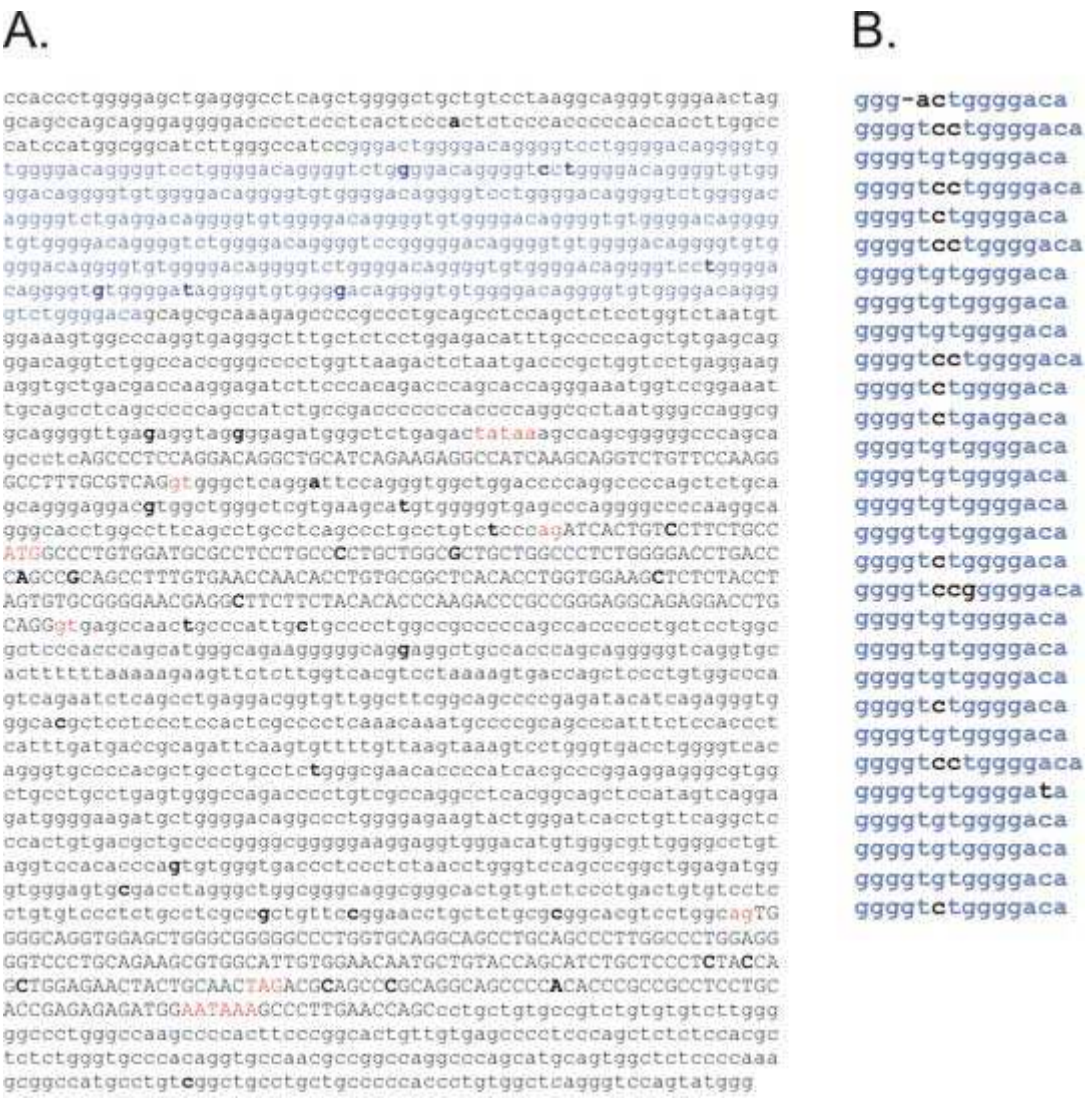


Fig. 8. Polymorphism in microsatellites can arise from polymerase slippage
During polymeration, thermal fluctuation can dissociate DNA strands. The reassociation is usually perfect, resulting in no change. However, occasionally, the DNA may align unequally, due to the repeat. Either the polymerizing strand loops back, which can result in expansion of the repeat (that is more frequent); or the elongating strand bind more distally to the template (the template loops back) with subsequent shrinkage of the repeat. Inset: Some repeats may promote this procedure due to stabilization of the transition state by forming stem-loop structure from the imperfect double helix, especially CAG/CTG repeat, which is involved in pathogenesis of several trinucleotide expansion diseases. The longer the microsatellite, the higher probability of polymerase slippage, which creates, in combination with the more pronounced tendention for elongation of the repeat, a positive (reinforcing) feedback loop.