U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Consequences of Alu-mediated recombination events

Dissertação de Mestrado em Genética Forense

ANA CAROLINA CARLOS TEIXEIRA DA SILVA

Faculdade de Ciências da Universidade do Porto

2012

CONSEQUENCES OF ALU-MEDIATED RECOMBINATION EVENTS

Dissertação submetida à Faculdade de Ciências da Universidade do Porto para obtenção do grau de Mestre em Genética Forense.

Dissertation submitted to the Faculty of Sciences of the University of Porto for the Master's degree in Forensic Genetics.

Instituição / Institution:

IPATIMUP

Instituto de Patologia e Imunologia Molecular da Universidade do Porto

Orientadora / Supervisor:

Doutora Luísa Azevedo

IPATIMUP

“Around here we don’t look backwards for very long…

We keep moving forward, opening up new doors and

Doing new things because we’re curious…

And curiosity keeps leading us down new paths”

Walt Disney

Table of Contents

Figures Index

Tables Index

Acknowledgements

This thesis was built with the help and support of many people; therefore I feel as I must thank all those who contributed to the success of this dissertation and/or influenced me to grow intellectually and personally during this past year.

My most special word of acknowledgement goes to my supervisor Luísa Azevedo, to whom I owe a great deal for guiding me far beyond the scientific matters, for encouraging me to be critical and creative in every aspect of this project, for motivating me to achieve my goals and for helping me discover what I truly like about biology. For all these reasons and far more, a huge thank you!

I also thank Professor António Amorim for the active interest and participation in this project, and constant availability to assist during the most challenging parts.

A special word of gratitude goes to the other co-authors of the article and/or posters of this project, Raquel Silva and João Carneiro, for all the help, feedback and critical reviews that were certainly vital to the accomplishment of the project.

Also, a very special thanks to my Forensic Genetics Masters' classmates and friends, Alexandre Almeida, Catarina Xavier, Filipa Melo, Lídia Birolo, Marisa Oliveira, Nuno Nogueira and Sofia Marques for making this year extremely fun, for all the friendship, and for all the genius scientific brainstorms. A huge thanks goes to Alexandre for all the inspiration, support, friendship and love that he gave me and for being the most special and amazing person in my life.

Big thanks also to Catarina Seabra and Inês Martins that, for five years, have been by my side and for being the greatest friends and housemates a person can have. To my other friends and colleagues in Aveiro (Ana do Carmo, Joana Formigal, Renato Pinho) that encouraged me to pursue this area.

I would also like to thank the rest of the Population Genetics group and Sequencing Services for all the good moments and sympathy, especially to Sara Pereira who has helped me a lot in the laboratory, always kindly and patiently.

At last, I would like to thank my family, that despite being geographically distant, always aided me morally (and financially), and to whom I owe who I am today.

Abstract

Alus are the most successful transposable elements found in the primate genome, occupying about 10% of its sequence. These elements are categorised into subfamilies according to their retrotransposition-competent source gene and several diagnostic positions. Alus hold several characteristics useful for forensic analyses and can be used for individual identification, DNA quantification and other non-human applications. Furthermore, due to their homology and abundance, Alus are prone to recombination that can result in genomic rearrangements of clinical and evolutionary significance. For instance, disease-causing rearrangements in the ornithine transcarbamylase gene (*OTC*), located in Xp21.1, are known to be Alu-mediated.

In this study, the role of recombination in the origin of novel Alu source genes was addressed along with the classification system, through the analysis of all known *consensus* sequences compiled from literature and related databases. Furthermore, the frequency and structural organisation of the Alu elements within the *OTC* gene was also analysed in order to correlate them with possible rearrangements in the gene. A total of six polymorphic indel markers within the non-coding region of the gene were selected and compiled into a PCR multiplex, with the purpose of studying the haplotypic structure of the European population and use that information as a supporting diagnostic technique.

From the analysis of the entire collection of Alu *consensus* sequences, recombination was identified as the origin of two particular subfamilies: AluSx4 and recent subfamilies of young Alus (Y). These results demonstrate that active Alus can arise from ectopic recombination and regain retrotransposition ability. Additionally, the results reveal a new potential use of Alus in forensic analyses as subfamily polymorphism, an area that could be further explored. Concerning the *OTC* gene, a whole gene scan revealed a total of 28 Alu elements. The distribution of these Alu elements between the sense and the antisense strand showed to be similar and widespread through the gene, revealing that ectopic recombination is expectedly frequent, and that the *a priori* probability of a deleterious rearrangement is equally distributed across the gene. This reinforces the fact that supporting diagnostic approaches are needed to detect such rearrangements. Patterns of *linkage disequilibrium* between the markers led us to consider the hypothesis of the presence of two recombination hotspots located in the low Alu density region of the gene. All these results have posed even more questions regarding the role of Alus in shaping the human genome, ultimately encouraging further research.

Resumo

Os Alus são os elementos transponíveis mais bem sucedidos no genoma dos primatas, ocupando 10% do seu conteúdo. Os Alus classificam-se em subfamílias de acordo com o gene-mestre que lhes deu origem e segundo as mutações diagnóstico que possuem. Estes retrotransposões possuem características de interesse para análises forenses, sendo utilizados na identificação indivual, quantificação de DNA e em análises de amostras não humanas. Devido à sua elevada homologia e abundância, os Alus têm tendência a recombinar, podendo estes eventos culminar em rearranjos genómicos de importância clínica e evolutiva. O gene da ornitina transcarbamilase (*OTC,*) localizado na região Xp21.1, é um dos exemplos de genes em que já foram descritos estes rearranjos deletérios mediados por Alus.

O tema central deste trabalho consistiu em estudar o papel da recombinação na origem de novas subfamílias de Alus. Além disso, procurou-se reavaliar o sistema de classificação de subfamilias atualmente usado, através do estudo de uma compilação de sequências *consensus* de Alus retiradas de bases de dados e da literatura. Adicionalmente, estudou-se o gene da *OTC* em relação ao seu conteúdo de Alus, de modo a tentar relacionar a sua densidade e distibuição com a ocorrência de possíveis rearranjos. Desenvolveu-se, também, um sistema de PCR-multiplex com base num conjunto de seis indels polimórficos, com o propósito de se estudar a estrutura haplotípica da população europeia e usar esta informação como suporte ao diagnóstio da deficicência de OTC.

Através da análise das sequências *consensus* de Alus, conseguiu-se detetar duas subfamílias que tiveram origem em eventos recombinacionais: a AluSx4 e uma família de Alus Y (não especificada). Estes resultados demonstram que os Alus ativos podem surgir por recombinação ectópica e voltar a ganhar capacidade de retrotransposição. Em adição, estes resultados revelaram uma potencial nova aplicação destes retrotransposões como polimorfismos de subfamília, no ramo forense, uma área que poderá ser explorada no futuro. Uma análise da sequência completa do gene revelou um total de 28 inserções de Alus. A sua distribuição pelo gene é equilibrada, indicando que a probabilidade de ocorrência *a priori* de um rearranjo deletério é igualmente distribuída pelo gene. A abordagem PCR-multiplex aqui desenvolvida e os estudos preliminares aos padrões de *linkage disequillibrium* do gene revelaram dois possíveis *hotspots* de recombinação dentro do gene, localizados em zonas com baixa densidade de Alus. O conjunto dos resultados obtidos neste estudo colocou ainda mais questões no que toca ao papel dos Alus na arquitetura do genoma humano, demonstrando a necessidade de prosseguir investigações futuras.

Abbreviations

A – Adenine

Array-CGH - Microarray-based comparative genomic hybridisation

ARMD – Alu recombination-mediated deletion

Bp – Base pair

C – Cytosine

cDNA – Complementary DNA

CpG – Cytosine-phospho-guanine

Del – Deletion

dHJ – Double HJ

DNA – Deoxyribonucleic acid

DSB – Double strand break

DSBR – Double strand break repair

FLAM – Free left Alu monomer

FRAM – Free right Alu monomer

G - Guanine

HERV – Human endogenous retrovirus

HJ – Holliday junction

Indel – insertion / deletion

Ins – Insertion

Kb – Kilobases

L1 – LINE-1

L2 – LINE-2

LINE – Long interspersed nuclear element

LTR – Long terminal repeat

MIR – Mammalian-wide interspersed repeat

MLPA – Multiplex ligation-dependent probe amplification

mRNA – Messenger RNA

Myr – Million years

NAHR – Non-allelic homologous recombination

ORF – Open reading frame

OTC – Ornithine transcarbamylase

OTCD – OTC deficiency

PCR – Polymerase chain reaction

PCR-SSCP – PCR- single strand conformation polymorphisms

RFLP – Restriction fragment length polymorphism

RNA – Ribonucleic acid

RNA pol III – RNA polymerase III

RNP – ribonucleoprotein

SINE – Short interspersed nuclear element

SNP – Single nucleotide polymorphism

SRP – Signal recognition particles

STR – Short tandem repeat

SVA – SINE VNTR Alu

T – Thymine

TE – Transposable element

TPRT – Target-prime reverse transcription

UTR – Untranslated region

VNTR – Variable number tandem repeat

Introduction

Transposable elements

Genomic repetitive DNA is presented in two forms: tandem, when the repeat motifs are adjacent to each other, or interspersed, when repeats are spread all across the genome [1]. Transposable Elements (TEs) or "jumping genes" are short pieces of DNA with the ability to move within the genome [2]. Consequently, they are represented by numerous dispersed copies (Figure 1), both in prokaryotes and eukaryotes [3]. In humans, they constitute up to half of the genome [4]. TEs are subdivided into two categories: DNA transposons and retrotransposons (Figure 1).

DNA transposons move by a "cut-and-paste" mechanism, i.e. they can cut and insert themselves into different parts of the genome. These elements account for ~3% of the human genome and are currently not mobile due to mutation accumulation [3].

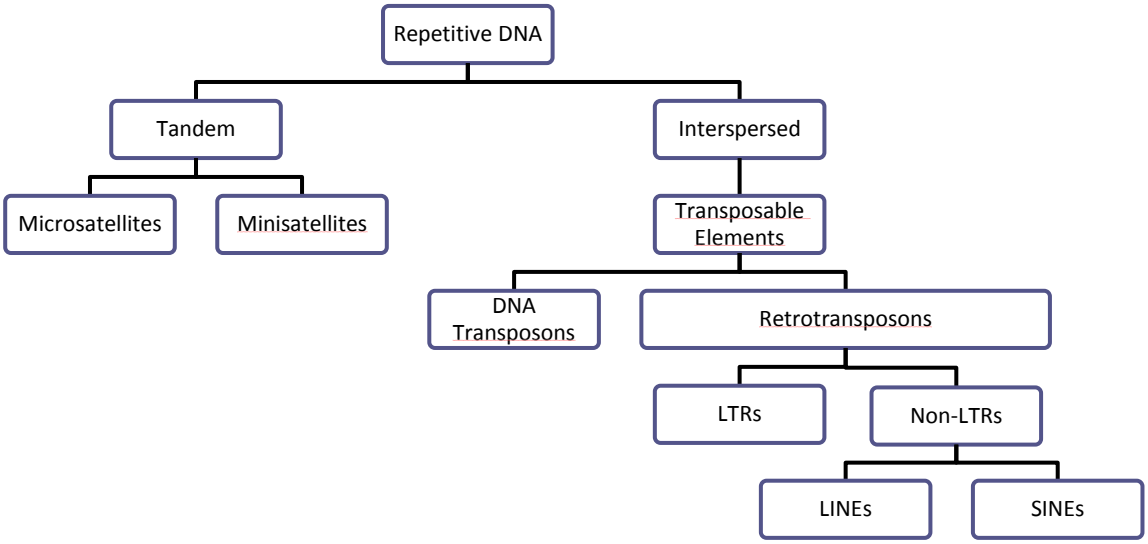


Figure 1: Organisation of repetitive DNA.

Retrotransposons, however, move by a "copy-and-paste" mechanism through RNA intermediates that are reverse transcribed and then inserted as cDNA copies in distinct locations [5, 6]. Retrotransposons are classified into two sub-groups, according to the presence or absence of Long Terminal Repeats (LTRs). LTRs are segments of 300 to 1000 base pairs (bp). In humans, they correspond to the Human Endogenous Retroviruses' (HERV) sequences and account for ~8% of the genome with little or none on-going activity, again, due to the accumulation of

inactivating mutations [4, 7]. Non-LTR retrotransposons are the major human components of TEs. This class includes the Long Interspersed Nuclear Elements (LINEs), whose most abundant elements are the LINE-1 or L1, and the Short Interspersed Nuclear Elements (SINEs) that include the SVA (SINE VNTR Alu) and the Alu elements. L1s, SVA and Alu elements are the only non-LTR elements with proven remaining retrotransposition ability [8, 9]. The other genomic non-LTR elements, such as LINE-2 and Mammalian-wide Interspersed Repeats (MIR), are inactive and only comprise ~6% of the genome [4].

L1 elements represent about 17% of the human genome with over half a million copies [4]. They are 6 Kb long and encode the necessary machinery for their own retrotransposition in their two open reading frames (ORF1 and ORF2) [10, 11], which makes them the only autonomous TE in the genome. The integration process is known as target-primed reverse transcription (TPRT). Nevertheless, not all of the resulting L1 copies are capable of being retrotransposed since many suffer truncation, rearrangements and impairing point mutations. In fact, only less than 100 L1 copies are currently known to be active [4, 12]. Active L1 elements also harbour the essential machinery for the dissemination of other active TEs: SVAs and Alus [6, 13], being thus responsible directly or indirectly for all the recent *de novo* TE insertions.

SVA elements are complex SINEs with approximately 2 Kb of length. They consist of a multipart structure involving an hexamer repeat region followed by an Alu-like monomer, a variable number tandem repeat (VNTR) region, a HERV-like region and a poly-adenine 3' tail [13, 14]. There are ~3000 copies of SVA elements in the genome; however, as mentioned above, none of them hold the necessary machinery for mobilisation. Instead, these elements take advantage of the L1 retrotransposition to move across the genome [13, 14], as do Alu elements.

TEs can cause mutations in the host genome either by insertion in new locations, when moving from one part of the genome to another or, in a post-insertion stage, by creating numerous regions with high homology and consequently promoting recombination between non-allelic DNA sections [3]. This mechanism was the core of this project, which mainly focused on the consequences of rearrangements caused by Alu elements, the most frequent class of SINEs.

Alu elements

Origin and Structure

The Alu family of retrotransposons is primate-specific, dating back to 65 million years (Myr) ago [15]. A common Alu element is about 300 bp long and is composed by two homologous monomers, left and right, with origin in the terminal segments of the signal recognition particle RNA, also known as 7SL RNA (Figure 2). These monomers are termed Free Left Alu Monomer (FLAM) and Free Right Alu Monomer (FRAM), respectively, when they are found loose in the genome. Connecting the monomers is an adenine-rich linker and another A-rich region flanks the 3´end of these elements [16].

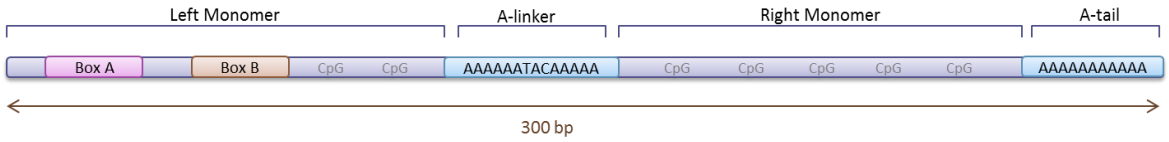


Figure 2: Alu structure.

The left unit is about 140 bp long [17, 18]. Within its sequence, there is a two-part internal promoter for the RNA polymerase III, located in boxes A and B [19]. Both these boxes are proximately 10 bp long [20] and they are located around positions 10 and 70, respectively [20-22]. The specific functions of boxes A and B are enhancing transcription and specifying the position of the transcription site upstream of box A [23]. Defects in these sequences are likely to impair the Alu retrotransposition ability. The right monomer is larger, containing 31 additional bases [17, 18], however it does not contain any promoter sequence and no specific function in Alu transcription is known.

A central A-rich sequence (linker) connects the monomers. The typical sequence is A_6TACA_5, still, as a mononucleotide microsatellite, strand slippage and point mutations make this a rather unstable region. The linker, along with the poly-A tail at the 3'end, is a source of origin and expansion of microsatellites [24].

The poly-A tail at the 3'end is responsible for priming the reverse transcript during the integration phase of retrotransposition (Figure 3). The tail is the most mutable region of the Alu,

yet its length and homogeneity are critical features for retrotransposition activity [25]. In that sense, Alus with tails longer than 40 bp and long stretches of pure adenines have higher chances of retrotransposition success [25]. A-tail retraction is observed in older Alus, as they tend to possess a shorter 3' A-stretch than younger ones. However, cases of A-tail expansion were discovered and associated with strand slippage [26] and unequal recombination (partial gene conversion of the A-tail). These alterations enable the resurrection of otherwise inactive Alus [27]. The accumulation of point mutations increases sequence heterogeneity, and can help to stabilize this region in terms of strand slippage or, result in microsatellite origin and expansion.

Distribution and abundance across the genome

As a result of their continuous mobilisation during the past 65 Myr [19], there are currently over a million Alu elements [4], comprising over 10% of the human genome. For this reason, they are considered the most successful transposable element in the human genome [28].

Like other SINEs, Alus mostly occupy non-coding domains of genes: introns, upstream and downstream flanking regions, and inter-genic areas [29]. This biased distribution towards gene-rich areas is unlikely the result of any type of insertional preference [30] , but rather a result of Alu depletion due to recombination-mediated deletion in gene-poor regions. These events in gene-rich areas are not likely to be inherited due to their often deleterious effects [19].

Retrotransposition

The process by which non-LTR elements spread through the genome is called retrotransposition, since this is a RNA-based copy number amplification [31, 32]. A cDNA molecule generated by the reverse transcription of the Alu RNA is inserted into a new location [32, 33].

As Alu elements have no coding capacity, they are classified as non-autonomous elements. They rely on the L1-encoded proteins for their own transposition [7]. In order to grasp the concept of Alu mobilisation, it is necessary to understand the LINE-1 retrotransposition mechanism. The first step of retrotransposition involves the transcription of an L1 *locus* by RNA polymerase II from an internal promoter that drives the transcription from the 5' end of the L1 element [10, 34]. In the cytoplasm, ORF1 and ORF2 are translated. These two ORFs encode an RNA-binding protein (ORF1), and a protein with endonuclease and reverse transcriptase

properties (ORF2). These proteins bind to the L1 RNA transcript to form a ribonucleoprotein (RNP), which is transported back into the nucleus to initiate the integration process [35].

The integration of the L1 occurs through a process called target-prime reverse transcription (TPRT) [35-37]. The endonuclease cleaves the first strand of targeted DNA between the T and the A of a specific sequence 5'-TTAAAA-3' [38]. The poly-A tail of the L1 RNA sequence pairs with the Ts of the host DNA, and a sequence complementary to the L1 RNA is generated. Occasionally, another strand of the host DNA is cleaved at a second nicking site with a less conserved sequence 5'-ANTNTNAA-3' located at a variable distance from the first nicking site [39]. The newly inserted fragment of single strand cDNA is used as template for the synthesis of the second strand of the L1 fragment. During this process, truncation of 5' segments and point mutations are frequent [4, 12].

On the other hand, Alu transcription is done by RNA polymerase III (Figure 3A). Alu transcripts travel to the cytoplasm and connect to the signal recognition particles (SRP) 9 or 14 to form RNPs (Figure 3B). Active Alu elements' integration seems to occur mainly by TPRT as well (Figure 3B-F), however, these elements need to highjack L1 machinery to do so [6]. The source of the reverse transcriptase for the generation of Alu cDNA from RNA is uncertain; though it is most likely provided by L1s [37, 40].

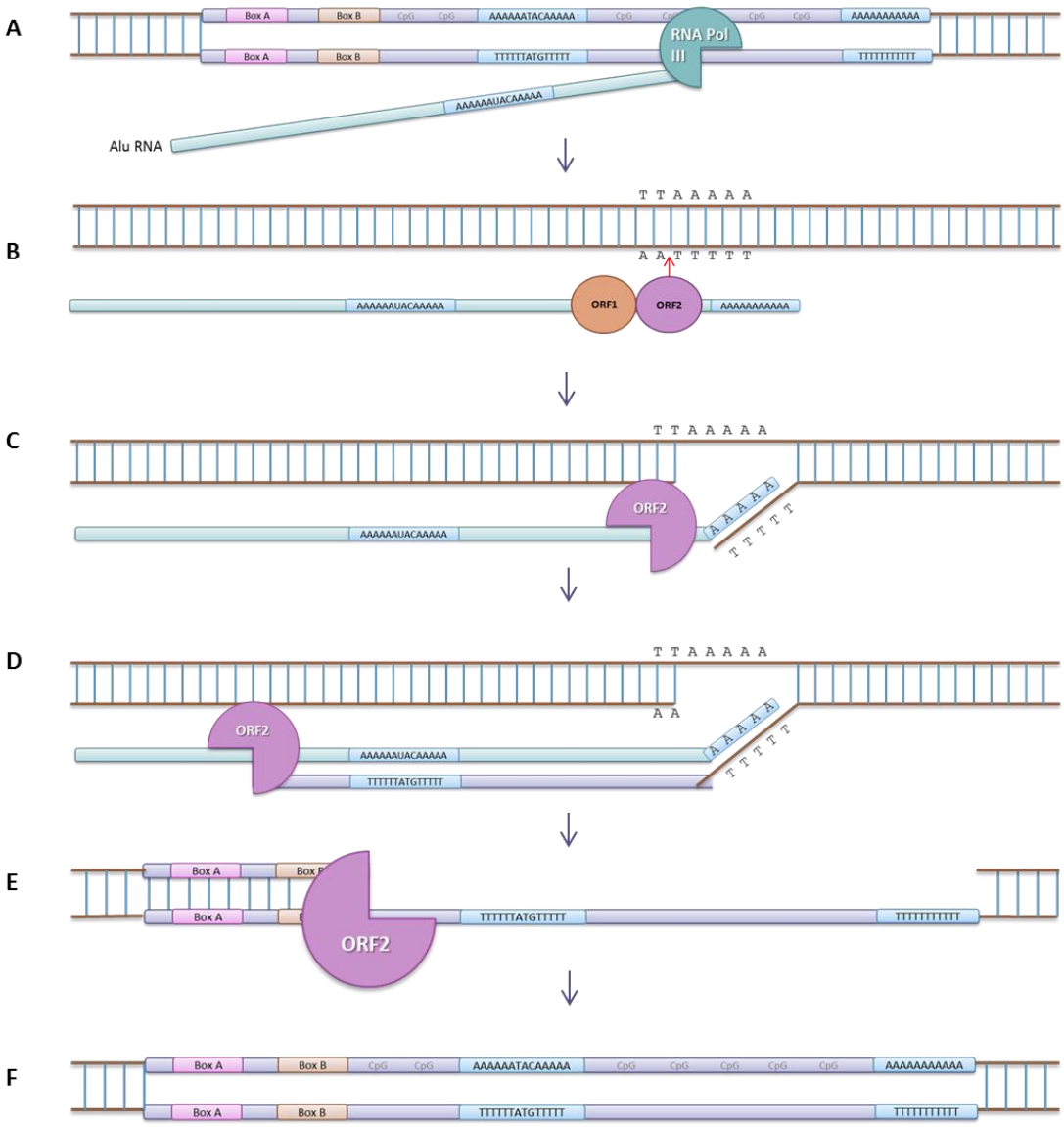


Figure 3: Alu retrotransposition. (A) Alu transcription by RNA pol III; (B) ribonucleoprotein formation and host DNA cut; (C) priming of the Alu RNA to the host DNA; (D) Alu cDNA synthesis; (E) second DNA strand synthesis; (F) completed retrotransposition.

Alu inactivation

Although the genome of primates is full of Alu copies, only few are capable of dissemination. Older Alu elements tend to be inactive, whereas some young ones may still hold retrotransposition ability. There are a number of possible causes for retrotransposition impairment, including transcriptional limitations or problems in Alu integration [41].

Point mutations or truncation in an Alu may result in loss of retrotransposition ability if the promoter sequence is affected [19, 42]. In a post-transcriptional stage, retrotransposition conclusion may be impaired due to instability of Alu RNA secondary structure, difficulties in ORFs - Alu RNA interactions [25] or difficulties in priming the Alu transcript.

Classification – Subfamilies

The categorisation of Alus into subfamilies is defined by specific alterations (diagnostic mutations) relatively to the original sequence that occurred during transpositional waves in the past 65 Myr. Hence, the establishment of a new subfamily is explained by the progressive accumulation of mutations relative to the parental subfamily [43]. This system of classification is useful to trace back the history of a transposon and to access the active/inactive status [14, 44].

The three major Alu subfamilies are the ancient AluJ, the intermediate AluS [45] and the young AluY. The retrotransposition activity of the AluJ subfamily dates back to at least 60 Myr, while the AluS had its main activity status between 60 and 20 Myr ago [46] and AluY in the past 20 Myr and some members are still active nowadays [47]. These tree major clusters are subdivided into other smaller subfamilies. Currently, 74 human subfamilies of Alus are known based on related databases and literature. Most of those are shared with other primates and a few (Yc1, Yc2, Ya5, Ya5a2, Ya8, Yb8 and Yb9) are human-specific [41, 48-58]. Altogether, there are about 2000 human-specific Alu elements, corresponding to only 0.5% of all Alus in the human genome [59].

Nomenclature

In order to unify new subfamily designations, nomenclature was standardised in 1996 by Batzer *et al* [60]. In this system, which is currently used, a capitalised letter indicates the major subfamily (J, S and Y), followed by a lowercase letter in alphabetical order, based on the order of publication, which indicates a sub-branch and the number of diagnostic mutations relative to the major subfamily.

Subfamily consensus sequences

The *consensus* sequence of a specific subfamily is the predicted sequence of the first (active) subfamily source-gene, even if it no longer exists its active form [61]. This way, mutations that are shared by Alus of the same subfamily also appear in the *consensus* sequence and are thus called diagnostic positions. The general *consensus* sequence does not correspond to the AluJ as it would be expected. Instead, since the AluSx subfamily is the most abundant in the human genome, it represents the general human Alu *consensus* sequence [62].

Source genes

The source, or master gene, of each subfamily is an active element with the ability to generate new Alu copies [63]. Currently all AluY and most of the AluS subfamilies possess active source genes, in contrast to older subfamilies such as AluJ. The number of source genes for each subfamily is very low, indicating that (a) most of the copies are inactive and, (b) that they originated from a very low number of source genes. Despite the fact that only a small percentage of Alu copies are active, they outnumber by far all other TE active copies in humans (reviewed in [7]).

Alu amplification rate

The human genome encompasses about 300 million recent insertions in addition to several million fixed TEs [4]. It is estimated that a new Alu insertion occurs every 20 live births [64], but this amplification rate has not been uniform over time. The majority of Alu insertions occurred about 40 Myr ago, reaching one insertion in every birth [43]. Nowadays, there seems to be a general tendency for relaxation of Alu retrotransposition, decreasing the impact of these TEs in the genome.

Alu-mediated genome shaping

Previous studies [65-67] have shown that Alu elements have had an important role in the evolution of the primate genome. Changes in the genome architecture by Alus, and TEs in general, are mainly due to insertion-mediated deletions [68, 69], and recombination mediated rearrangements such as deletions [70, 71], segmental duplications [72, 73], inversions [74] and translocations [75, 76].

De novo Alu insertion consequences

The most obvious consequence of a continuous retrotransposition activity of Alu elements is the increase of genome size [77]. Paradoxically, Alu insertions may also cause deletions (Figure 4), thus diminishing the effect of genome size extension. Insertions of Alu elements results in the deletion, by endonuclease dependent or independent mechanisms, of a portion of adjacent sequence occasionally larger

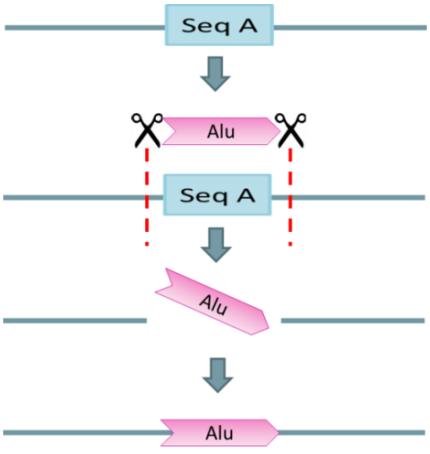


Figure 4: Alu insertion-mediated deletion.

than the Alu insert itself [68].

Another consequence of this enduring process is the creation of inter-individual variation of Alu copy-number [78, 79]. These polymorphic Alu insertions (presence or absence) are very useful genetic markers for evolution, demography and forensic studies [80-82].

Alus can also alter the architecture of a gene upon insertion into coding or regulatory regions. Depending on the insertion location and the affected gene, this process may have deleterious effects [8, 59]. It is estimated that about 0.1% of all human genetic disorders are generated by this process [59].

Double strand breaks (DSBs) are directly associated with L1 ORF2 endonuclease activity [83], which is critical for both L1 and Alu insertions. However, the number of DSBs is much higher than the actual TE insertion. DNA DSBs are one of the most lethal types of DNA damage. A DSB can on its own kill a cell or disrupt its genomic stability [84]. On the other hand, Alu elements and other non-LTR elements can also act as containment measures against DSBs because they can invade and repair the cleaved sequence [85].

There are evidences that Alu insertions have other effects in the human genome. By means of several different mechanisms, such as modulation of gene expression, RNA editing, epigenetic regulation and conservation of non-coding elements, they are able to control gene expression (topics reviewed by [65]). Alus are as well associated with the emergence of orphan genes and exonisation processes, due to the fact that they contain motifs that can become functional splice sites via specific mutations [86], generating functional protein variants [87].

Recombination

The recombination process allows the exchange of sections between molecules of DNA [88], based on sequence homology of the segments involved during mitosis and meiosis. Meiotic recombination occurs during prophase I, with the pairing of homologs. This pairing is dependent of the homology between DNA strands and is considered to be a transitory and unstable connection [89, 90]. Several models for this process have been described; yet, the most accepted is the double-stranded DNA break repair model (DSBR). According to this, recombination starts with a DSB on one of the molecules, followed by 5' strands retraction, generating 3' single-stranded extremities. One of these 3' extremities infiltrates into the other molecule using its sequence as a template for DNA synthesis. Then, a double Holliday junction is

formed and its configuration determines if the recombination type is crossover or gene conversion (Figure 5) [88, 91, 92].

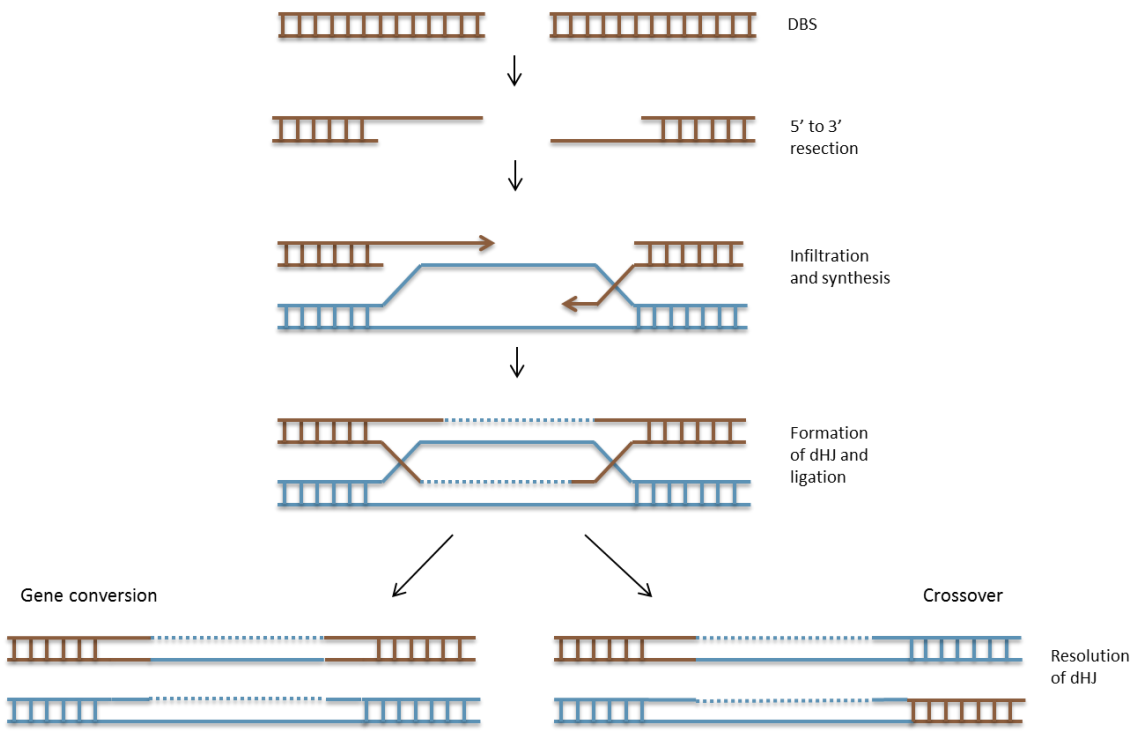


Figure 5: Recombination: gene conversion and crossover.

In most cases, recombination does not create structural variations. However, when recombination occurs out of the homologous locations (ectopic recombination), genomic rearrangements can arise [71], which may cause phenotypic changes [9, 59].

At a post-insertion stage, Alu elements continue to shape the primates' genomes through the process of recombination [93], by means of crossing-over and gene conversion. Due to their proximity in the genome (one insertion every 3 Kb), high GC content (~62.7%) and high sequence similarity (70%-100%) Alus are prone to successful recombination [19, 59]. Alu-mediated recombination events can occur in the somatic or in the germ line [19].

It is currently acknowledged that there is a positive correlation between sequence identity and recombination events [71]. Alu elements have equal probability of recombining, regardless of the subfamily they belong. These observations can seem rather contradictory, since elements from the same subfamily should have higher sequence identity (and therefore a

higher probability of recombining) than different subfamily members. Nevertheless, this is easily explained by the existence of numerous truncated Alu elements that result in lower identities between members of the same subfamily when compared with members of different subfamilies that remain intact. Thus, the principal effects of Alu-Alu mediated rearrangements were observed in early primate evolution when a higher proportion of Alu elements were more identical to one another [59]. Interestingly, there are studies [95] that point to Alu insertions reducing recombination events in its neighbourhood. During early primate evolution, this preclusion of chromosomal recombination may possibly have aid speciation, via chromosomal incompatibility [19].

Crossover is a reciprocal trade of homologous segments in which both chromosomes exchange a portion with the other. This type of recombination is of extreme importance for meiosis, allowing the correct segregation of chromosomes [96, 97]. Despite that, crossover is the least common resolution of recombination (less than 8%) [98], so most of the DNA sequence shuffling is the result of gene conversion.

Gene conversion is a type of recombination characterised by the non-reciprocal transfer of homologous DNA sequences from a donor to an acceptor. This process is initiated with a DSB, either caused by the enzyme SPO11 during meiosis or by other factors (radiation, stalled replication forks, etc) in mitosis. During its course, genetic information is transferred from a homologous region (donor) to the region that contains DSB (acceptor) [99, 100]. There are currently three models of gene conversion: the seminal double strand break repair, the synthesis-dependent strand-annealing and the double-HJ (Holliday Junction[1]) dissolution reviewed in Chen *et al* 2007 [101]. Gene conversion itself seldom culminates in genomic rearrangements [102].

These events can occur between non-alleles (non-allelic gene conversion) or between alleles (inter-allelic gene conversion). Nearly all cases of deleterious gene conversion are due to non-allelic events, particularly within the same chromosome (intra-chromosomal). In contrast, the occurrence of inter-allelic events seldom causes genetic diseases. Non-allelic gene conversion also has consequences to concerted evolution[2], as so paralogous sequences become more closely related to each other than to their orthologous. Sequence homogenisation due to gene conversion increases the likelihood of non-allelic recombination by increasing the number

[1] Holliday Junction is the location in which two DNA strands exchange sequences during recombination.

[2] Concerted evolution designates a process of homogenisation of repetitive DNA family between individuals of the same species, such that they become more closely related between themselves than they do with their orthologous in other species.

of sites with high homology, contributing to genomic rearrangements in an indirect form [103, 104].

Gene conversion events usually require a sequence homology of over 92% [101], and the rate of gene conversion is directly proportional to the length of identical bases [105, 106]. In mammals, gene conversion tracts[3] tend to range from 200 bp to 1 kb. Regardless of their short size, Alu elements frequently undergo gene conversion [102, 107] because they present high values of identity between them.

Detecting gene conversion events is extremely important because Alu gene conversion acts as a secondary pathway for Alu mobilisation within the genome, further increasing Alu homology sites, and facilitates genomic rearrangements through sequence homogenisation (concerted evolution) [71]. However, it is also involved in sequence variability, via partial gene conversion between Alus from different subfamilies. This way, gene conversion contributes to inter-subfamily differences, inactivation or re-activation of Alus by partially converting non-functional or functional portions (respectively) from an Alu to another [19].

These phenomena are difficult to be proved in humans because the analysis of both products of a single recombination is impossible *in vivo* [101]. In addition, detecting Alu gene conversion is difficult because Alu elements are so closely related to each other that changes in their sequence caused by gene conversion are often masked as random point mutations [108]. Furthermore, events of gene conversion can only be distinguished from double crossover by the length of the converted tract, since it is considerably larger in double crossovers[4].

Ectopic recombination and genomic rearrangements

Meiotic recombination normally occurs between alleles in homologous chromosomes. Nevertheless, due to the existence of high similarity regions dispersed throughout the genome, this mechanism can also happen between non-allelic, yet homologous, segments, such as Alu elements. These events are named non-allelic homologous recombination (NAHR) or simply ectopic recombination. In fact, NAHR can take place between homologous and non-homologous chromosomes (inter-chromosomal recombination), or even within the same chromosome (intra-

[3] Gene conversion tracts correspond to the donor sequence transferred to the acceptor. Its length is indicated in terms of minimum and maximum length, due to the impossibility to precise the breakpoints.
[4] Double crossover refers to two crossover events that result in the reciprocal transfer of an internal portion (or two external) of the chromosome. This transferred tract has a larger length than the ones originated from gene conversion.

chromosomal). As a consequence of these defective chromosomal joints, genomic rearrangements such as deletions, duplications and inversions can emerge [71].

Alu Recombination-mediated deletions (ARMDs) cause an even higher number of human genetic disorders than Alu *de novo* insertions [59]. Altogether, NAHR is responsible for about 0.3% of human genetic disorders [59], and accounts for 22% of the bulk of germline structural variation [109]. NAHR occurs at a rate of one event every 300 meioses, or 10^{-9} to 10^{-8} per generation [110]. Genomic rearrangements generated by ectopic recombination include deletions, duplications and inversions.

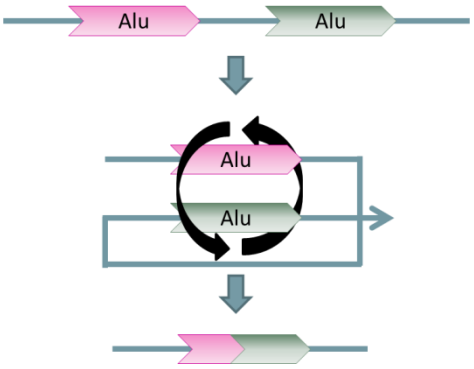


Figure 6: Alu-mediated intra-chromosomal recombination between Alus in the same sense resulting in sequence deletion and Alu chimerisation.

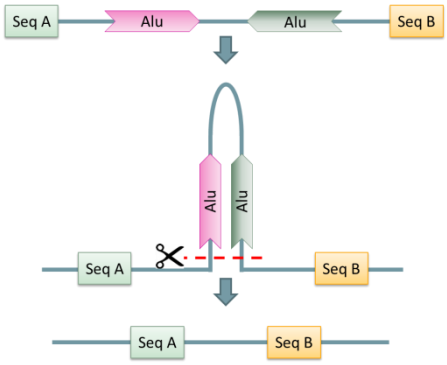


Figure 7: Alu-mediated intra-chromosomal recombination between Alus in opposite senses resulting in hairpin formation and excision.

ARMDs decrease the genome size by several mechanisms including intra- and inter-chromosomal recombination. These deletions usually produce chimeric and uninterrupted Alu elements (Figures 6 and 8) [71]. These deletions have an average size of 800 bp, but can range from ~100 to ~7300 bp and, since they occur in gene-rich regions, it is not surprising that over 70 reported cases of ARMDs account for numerous genetic disorders [9, 59]. In addition comparative genomics approaches unveiled almost 500 ARMD events since the human-chimpanzee divergence, underlining their species-specific effect in evolution [71].

The human genome encloses large segmental duplications (Figure 8), whose boundaries are Alu-rich, suggesting these elements had an important role in such rearrangements [72]. Alu-mediated recombination duplications contribute to the increase of the genome size, simultaneously increasing the number of high homology sites, and stimulating further recombination.

Comparative genomic approaches have been used to explore the contribution of Alu elements to chromosomal inversions (Figure 9). About half of the inversions that occurred in the human and chimpanzee genomes are retrotransposons-mediated. Despite the fact that this type of rearrangement does not involve gain or loss of genetic material, it has an important role in creating genomic variation and, in some cases, with functional consequences [111].

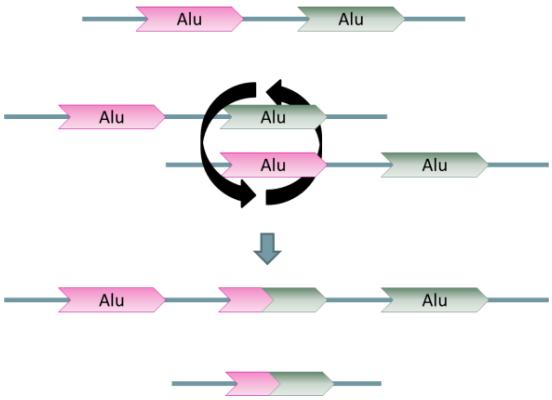


Figure 8: Alu-mediated inter-chromosomal recombination, resulting segmental duplications or deletions, and Alu chimerisation.

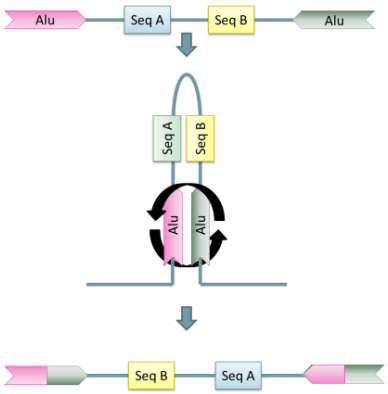


Figure 9: Alu-mediated intra-chromosomal recombination, resulting in sequence inversion and Alu chimerisation.

The role of recombination, namely gene conversion, as a source of Alu variability is a growing study-target. Studies on subfamilies AluYa [112] and AluYg6 [113] revealed that some of their elements possess intra-subfamily heterogeneity due to gene conversion that produced the chimeric sequences. Furthermore, genomic comparisons between orthologous *loci* in humans and other primates revealed, within the same *locus*, insertions of elements from different subfamilies as a result of gene conversion [114]. Moreover, the ability to regain retrotransposition-competence by restoring a functional poly-A tail, has been also attributed to gene conversion [27].

Microsatellite expansion

Due to their high copy number and structure, Alu elements can generate microsatellites or short tandem repeats (STRs) in the genome. These elements possess two regions that can undergo mutations, potentially generating new microsatellites: the middle A-rich linker and the 3' poly-A tail [24, 115]. About 20% of all microsatellites shared by humans and chimpanzees are located within Alus, including 50% of mononucleotide STRs [116]. There are some published

examples of Alu-mediated STR expansion that led to genetic disorders [117, 118], but most of these Alu-generated microsatellites are not deleterious.

Alu as genetic markers

Phylogenetic markers and taxonomic applications

SINE insertion polymorphisms are useful in phylogenetic analyses [119] because, once inserted, these are very stable markers, without relapse [81, 120], and with extremely low probability of independent insertions in the exact same location [59]. Since these elements are only present in primates' genomes, this type of analyses is only possible within this taxon. There have been a number of questions resolved using Alu elements, such as the human-chimpanzee-gorilla trichotomy [121] and the branching order of families of New World primates [122]. In these studies, the ability to target species-specific Alu subfamilies is of great importance. As a consequence of the sequential accumulation of Alus in the genome, a specific subfamily insertion can be correlated with a specific evolutionary period [123].

Forensic applications

Human genetic identification based on 32 polymorphic Alu insertions

At the present time, human genetic identification is based mainly in two types of genetic markers: the multiallelic markers STRs and the biallelic makers SNPs (Single Nucleotide Polymorphisms) [124, 125]. The use of both these marker types carries a two-step approach: (i) an initial PCR amplification and (ii) allele identification. This second step may be accomplished by several different methodologies that are usually expensive [126-128].

The human genome project came to reveal new potential genetic markers, the retroelements [4], with interesting features to human genetic identification purposes such as stability, neglecting probability of independent re-insertion in the same *locus*, and their simple identification [19, 129]. The main advantages in detecting these markers are the simplicity and the low cost involved [80], since it only requires a *locus*-specific PCR and agarose gel electrophoresis for detection.

Among all the families of retroelements, Alu elements are the most informative due to their high abundance and small size. Because they are recent insertions, the AluY subfamily elements are often used in these studies [80].

A total of 32 Alu insertion polymorphisms are currently used as human markers [80]: 31 of these in autosomes and one in the X chromosome for gender determination. This type of marker has been gaining increased acceptance among geneticists.

Quantification of human DNA samples based on fixed Alu elements

DNA quantification in a sample is an essential step in forensic analyses, as this can determine the appropriate type of marker to be analysed [130]. For this purpose highly sensitive methods for human DNA quantification [131-136] have been developed based on the large number of fixed Alu elements.

The ornithine transcarbamylase gene (*OTC*)

One of the genes that is documented as having suffered Alu-mediated genomic rearrangements is the *OTC* gene [137]. In this project, the Alu content of this gene was analysed in order to better understand some of the mechanisms behind the rearrangement-associated OTC deficiency. The *OTC* gene encodes the second enzyme of the urea cycle [138], and is mostly expressed in the liver and intestinal mucosa [139]. It is located in the short arm of the chromosome X, in Xp21.1 [140], and is organised in ten small exons and nine introns (Figure 10) [141].

Figure 10: Structural scheme of the *OTC* gene; exons are coloured blue, introns are coloured green and 5' and 3' UTRs are coloured purple.

OTC deficiency (OTCD)

OTC deficiency (OTCD, MIM 300461) is the most common urea cycle disorder [142].

The OTCD phenotype is caused by the deficiency of the mitochondrial enzyme ornithine transcarbamylase, a catalyser of the conversion of ornithine and carbamyl phosphate into citrulline [143], involved in the second step of the urea cycle [140]. As a consequence of the impairment of the urea cycle, patients with OTCD show hyperammonemia [144]. Other biochemical manifestations of this disease include high blood levels of glutamine, low blood levels of citrulline, and increased excretion of orotic acid [145, 146].

Ornithine transcarbamylase deficiency is a semi-dominant trait [140]. A variety of mutations can cause OTC deficiency [147], producing a broad-spectrum of symptoms. The majority of disease-causing mutations in this gene are single nucleotide polymorphisms [138], however, large rearrangements also occur and are lethal in males. Recurrent mutational events are extremely rare and most of the mutations tend to be family-specific [148].

Types, symptomatology, prognostic and treatment

OTCD has heterogeneous clinical manifestations [142], depending on the gender of the patient and the severity of the clinical manifestations: early or late onset.

Since the *OTC* gene is located on the X chromosome, hemizygous males tend to present a severe phenotype [149]. Whenever there is a total impairment in the expression or function of OTC, the disease is lethal at birth. Females, on the other hand, due to random patterns of X-chromosome lyonisation in hepatocytes, show a wider range of phenotypic heterogeneity [150] which includes the total absence of clinical manifestations, a milder phenotype manageable with diet and medication, and death in the most severe cases.

Early onset OTCD constitutes a more serious and often fatal disease type [151]. In this case, symptoms include hyperammonemia, lethargy and coma and are detected in the first hours after birth. This type of OTCD is either fatal or causes severe brain damage [138]. There is no cure, but the symptoms can in some cases be controlled depending on the mutation type and its effect in the mRNA or protein.

Some affected individuals remain asymptomatic until adulthood, being classified as late onset OTCD patients. In these cases, symptoms are usually triggered by environmental factors, namely protein rich diets, infections or stress. The manifestations include migraines, vomiting, lethargy, confusion, ataxia, hypotonia, among others [152]. This type can be more easily controlled with medication and diet.

Treatment for OTCD consists in the adoption of a low protein diet combined with supplements of arginine, sodium benzoate and phenylbutyrate to remove excess of nitrogen [153], but in some cases liver transplant is necessary.

Genetic tests

Enzymatic diagnostic approaches for the OTCD, although effective, are extremely invasive. Since ornithine transcarbamylase is mainly expressed in the liver and the intestinal mucosa, enzymatic diagnostics for confirmation of OTCD involves liver biopsy. The risks involved in a liver biopsy, especially if performed in a fetus for prenatal diagnosis, outweigh its efficiency.

Several methods have been described as an alternative to traditional enzymatic diagnostic tools for the detection of the disease, including prenatal [154-161] and preimplantation [162] techniques. These methods are based on Southern blot analysis [158], RFLPs (Restriction Fragment Length Polymorphisms) [155, 160-163] and PCR-SSCP (single strand conformation polymorphisms) for the detection of the mutated exons or the exon/intron boundary of the *OTC* gene [164]. Presently, OTCD detection is based mainly on the screening of exons and intro-exon boundaries [165], the analysis of mRNA transcripts [166], multiplex ligation-dependent probe amplification (MLPA) [137, 167], oligonucleotide arrays-CGH [167-169], high-density single-nucleotide array [170] and *linkage disequilibrium* analyses [171].

Genomic DNA tests using peripheral blood are the first diagnostic step and consist on the amplification of all ten exons and exon-intron boundaries, followed by the screening of mutations by automatic sequencing [165]. Still, this approach fails to detect deep intronic and regulatory mutations [172], or large deletions in heterozygous females. In these cases, the analysis of liver *OTC* mRNA transcripts, followed by synthesis of cDNA and its subsequent analysis have revealed to be very effective [166]. However, because *OTC* is mainly expressed in the liver and the small intestine this approach is invasive and the analysis of the mRNA transcripts might be limited by the degradation of abnormal mRNA resulting in false negative results [166].

Large genomic rearrangements leading to OTCD can be detected using MLPA [137, 167], oligonucleotide array CGH [167-169], high-density single-nucleotide array [167-169] and *linkage disequilibrium* [171]. These techniques help identify most of the cases undetected by exon and exon-intron boundaries screening.

Purpose

This project focused on a broad-spectrum of contents ranging from the general study of Alu elements, to the design of a potential auxiliary diagnostic technique to detect large rearrangements within the *OTC* gene. The specific goals of this study were to:

- Construct a database of all polymorphic sites of Alu subfamily *consensus* sequences
- Investigate the evolution of Alu subfamilies
- Explore the role of recombination in subfamily evolution
- Review the current classification system of Alu elements
- Locate and classify *OTC* Alus
- Correlate potential normal and abnormal recombination sites within the *OTC* gene with the position of *OTC* Alus
- Identify neutral polymorphic indel markers in the non-coding region of the *OTC* gene and design a multiplex-based auxiliary diagnostic system to detect large rearrangements

Materials and Methods

Evolutionary history of Alu subfamilies

The detailed information on the retrieval of all known Alu *consensus* sequences and subsequent sequence comparison, construction of a database of Alu polymorphic sites, network assembly and inference of Alu subfamily evolutionary history are in the journal article manuscript entitled "The role of recombination in the emergence of novel subfamilies" presented in the "Results and Discussion" chapter (Section I).

Location and classification of *OTC* Alus

The reference sequence for the human *OTC* gene was extracted from the Ensembl [173] database (ENSG00000036473), and Alu elements within were scanned using the programs Repeat Masker [174] and CENSOR server [175]. Alignments and values of pairwise identity were obtained using the software Geneious [176]. Alus were classified by the Repeat Masker [174], CENSOR [175] and CAlu (http://clustbu.cc.emory.edu/calu/index.cgi) programs.

Multiplex design for the detection of *OTC* rearrangements.

Markers selection and validation

The types of markers selected for this study were biallelic insertion/deletion polymorphisms also known as indels. Indels were our primal choice due to their stability and low mutation rate.

Several neutral indel markers (Figure 11) were selected from non-coding regions (introns, 5' and 3' UTR) of the human *OTC* gene sequence of the Ensembl database (ENSG00000036473). Primers for all these pre-selected indels were designed with the assistance of the bioinformatic tools Primer3 [177], OligoCalc [178] and BLAST [179], avoiding polymorphic sites annotated in the Ensemble reference sequence. *In silico* analyses of all primer pairs revealed no primer dimers or hairpin formation, nor primer binding-sites polymorphisms.

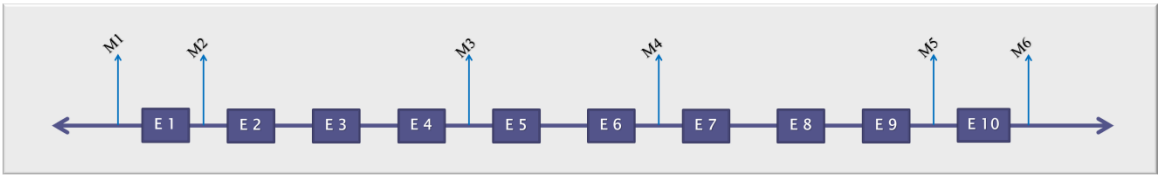


Figure 11: Relative location of the six indel markers analysed in the PCR multiplex

From those pre-selected markers, only six revealed to possess the desirable features for a successful multiplex design: their location across the *OTC* gene and their balanced allelic frequencies in the Caucasian European population (Table 1). The validation process was performed using a PCR singleplex and fragment sequencing[5]. Information relative to the markers, allele sizes and frequencies, and primer sequences are specified in Table 1 and Figure 12.

Table 1: Markers characteristics and primer sequences

Marker	Alleles	Size	Frequencies	Location		Primers sequence	Dye
M1	$(TTCT)_1$	232	0.78 (*n*=85)	24638	F	AAGGGAGCTCCAGGACTGA	FAM
	$(TTCT)_2$	236	0.22 (*n*=85)		R	GCTGCTGTGAAGGTGAGTA	
M2	$(AACTTA)_1$	211	0.25 (*n*=64)	26895	F	CCATTACACTGAGTTACATCAG	HEX
	$(AACTTA)_2$	217	0.75 (*n*=64)		R	TCAACTGTTTGGAGGAGGTTTT	
M3	$(ATACTT)_1$	200	0.27 (*n*=64)	62291	F	GCAGTGTACCAGAGCGTCAA	FAM
	$(ATACTT)_2$	206	0.73(*n*=64)		R	TGCGTGTGTCCTTTACAAGC	
M4	Del T	153	0.29 (*n*=56)	74744	F	GAGATCCATGCAGAGAAGATGA	FAM
	Ins T	154	0.71 (*n*=56)		R	AGGACAGCTCATTTTCCCTC	
M5	T_7	213	0.60 (*n*=62)	84589	F	GGTTCCAACTTGGTCATTCA	FAM
	T_8	214	0.40 (*n*=62)		R	CGGATCAAGGGTGGTAAGA	
M6	Del TG	183	0.44 (*n*=62)	106575	F	TTGTGCAGTGGGGAGTATTT	HEX
	Ins TG	185	0.56 (*n*=62)		R	GCAGTTCAGTTGAAGCGATG	

Multiplex optimization

All six markers were included into one single PCR multiplex reaction. Primers for these markers were marked with fluorescent dyes, allowing the simultaneous identification of all alleles by capillary electrophoresis. The optimized concentrations and volumes of the reagents used in this PCR are summarised in Table 2 and the PCR program is described in Figure 12.

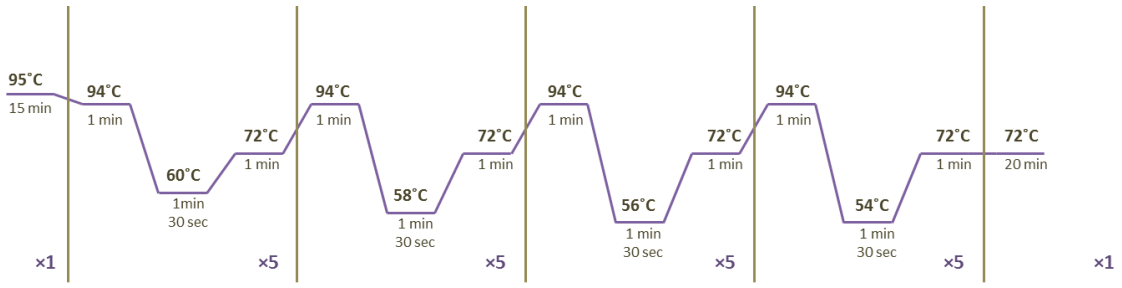


Figure 12: PCR multiplex program

5 These techniques include, after the first PCR reaction, an initial purification using ExoSAP-IT, to remove excess of primers and non-incorporated nucleotides, and a second purification using Sephadex after the sequencing reaction.

Table 2: Components of the PCR multiplex

Reagents	µL per tube		Concentrations
Qiagen Multiplex Master Mix	5		2×
H_2O	3		
Primer 1 F	0.07	0.5	2 µM
Primer 2 F	0.1		
Primer 3 F	0.07		
Primer 4 F	0.1		
Primer 5 F	0.1		
Primer 6 F	0.06		
Primer 1 R	0.07	0.5	2 µM
Primer 2 R	0.1		
Primer 3 R	0.07		
Primer 4 R	0.1		
Primer 5 R	0.1		
Primer 6 R	0.06		
DNA Sample	2		
Total	10		

In all PCR reactions, negative controls to detect possible DNA contaminations were used and amplification was confirmed by polyacrylamide electrophoresis with typical silver-staining procedures. Samples used are from anonymous blood donors and from a commercial DNA panel.

Fragment analysis

To 0.5 µl of PCR product were added 10 µl mix of formamide and ROX 500 (size marker). Fragment separation and sizing were performed by capillary electrophoresis in ABI PRISM 3130 Genetic Analyzer (from Applied Biosystems). Results were analysed in software Gene Mapper v4.0 (Applied Biosystems).

Results and Discussion

The results obtained in this work are presented in two sections as follows:

Section I: Data resulting from the analyses of Alu *consensus* sequence were compiled into a manuscript entitled "The role of recombination in the emergence of novel Alu subfamilies" which is presented in this section.

Section II: Data resulting from the study of the *OTC* gene in terms of Alu content and indel haplotypes

SECTION I

THE ROLE OF RECOMBINATION IN THE EMERGENCE OF NOVEL ALU SUBFAMILIES

Ana Teixeira-Silva[1,2], Raquel M. Silva[1], João Carneiro[1,2], António Amorim[1,2], Luisa Azevedo[1]*

1IPATIMUP-Institute of Molecular Pathology and Immunology of the University of Porto, Porto, Portugal

2 FCUP - Faculty of Sciences, University of Porto, Porto, Portugal

* Corresponding author: Luisa Azevedo, PhD., IPATIMUP, Institute of Molecular Pathology and Immunology of the University of Porto, Rua Dr Roberto Frias, S/N

4200-465 Porto, Portugal.

Telephone number: 351225570700

Fax number: 351225570799

Email: lazevedo@ipatimup.pt

Keywords: Transposable elements, Alu master gene, Alu subfamily, recombination, genome evolution

ABSTRACT

Alu elements are the most abundant and successful short interspersed nuclear elements found in mammalian genomes. In humans, Alus represent about 10% of the genome although less than 0.05% is active, that is, with retrotransposition ability. These elements are clustered into subfamilies of elements that evolved from the same retrotransposition-competent source gene(s). Alus are prone to recombination that can result in genomic rearrangements of clinical significance but have also an important role in the evolution of genomic structure. In this study, the role of recombination in the origin of novel Alu source genes was addressed by the analysis of all known *consensus* sequences of subfamily-specific source genes compiled from literature and related databases. From the allelic diversity analysis of the entire collection of Alu *consensus* sequences, distinct events of recombination were detected in the origin of particular subfamilies of AluS and AluY source genes. These results demonstrate that novel source genes can arise from ectopic recombination and strength the possibility that these chimeric elements can regain retrotransposition ability before proliferating throughout the genome.

INTRODUCTION

Alu elements are the most abundant and successful Short Interspersed Nuclear Elements (SINEs). These elements are exclusively found in primate genomes. In humans, they represent nearly 10% of the nuclear genome, that is, over 1 million copies and a frequency of one insertion per 3 Kb (Lander et al. 2001; Ullu and Tschudi 1984). An Alu is about 300 bp long and is composed by two monomers with origin in the 7SL RNA gene (Ullu and Tschudi 1984) attached one another by a poly-A stretch and punctuated by several CpG doublets. A second poly-A tail is present at the 3´end. Active Alus are those that intersperse the genome by retrotransposition, i.e. a cDNA molecule generated by reverse transcription of an Alu RNA is inserted in a distinct location (Rogers 1985; Weiner et al. 1986). Most of the Alus observed in a genome are relics of once active elements, as retrotransposition ability is often impaired by truncation of 5´ bases, shortening of the poly-A tail, or other mutations that occur during genome integration (Comeaux et al. 2009). Active Alu elements are accordingly called source or master genes.

Alu elements started to be classified in distinct subfamilies that diverged in specific (diagnostic) positions (Willard et al. 1987). Because events of back mutation and recombination, namely gene conversion (Zhi 2007), are frequent, such definition was later proposed to be changed to a collection of Alus that, at the moment of genomic integration, had origin in the same source gene (Styles and Brookfield 2007), though multiple source genes can contribute to an Alu subfamily (Matera et al. 1990)

Due to their proximity in the genome, high GC content (more than 60%) and sequence similarity (70%-100% of identity), Alus are prone to recombination (Batzer and Deininger 2002; Deininger and Batzer 1999) and a 13-mer DNA motif associated with recombination hotspots (CCNCCNTNNCCNC) is embedded in the sequence of some Alu subfamilies (McVean 2010;

Myers et al. 2002). Recombination between Alu sequences may lead to genomic rearrangements such as deletions, inversions and duplications that are of deleterious effect whenever gene-coding sequences are involved (Batzer and Deininger 2002; Deininger and Batzer 1999). Lynch Syndrome (Kuiper et al. 2011), OTC deficiency (Quental et al. 2009), Fabry Disease (Dobrovolny et al. 2011), hereditary spastic paraplegias (Conceicao Pereira et al. 2012) and some cancers are proven examples of Alu-mediated deleterious rearrangements (Batzer and Deininger 2002; Deininger and Batzer 1999). On the other hand, Alu-mediated rearrangements are as well believed to have had an important role in the evolution of primate genome (Han et al. 2007; Stoneking et al. 1997).

Gene conversion is assumedly critical in the evolution and spread of Alus (Zhi 2007). Previous data on specific subfamilies, for instances AluYa (Roy et al. 2000), and Yg6 (Styles and Brookfield 2007), genomic comparisons between orthologous *loci* in humans and other primates (Roy-Engel et al. 2002), and the ability to regain retrotransposition-competence by restoring a functional polyA tail (Johanning et al. 2003) motivated the search for the role of recombination in the origin of novel master genes contributing, this way, to the origin of novel Alu subfamilies. To answer this question, data mining for all known Alu *consensus* sequences was performed. Subsequent sequence comparison based both on single-nucleotide polymorphisms (SNPs) and insertion/deletion (indel) markers clearly revealed two cases of recombination: (a) between AluSq4 and AluSx3 resulting in the AluSx4 and, (b) between two unspecified elements that gave rise to either the cluster of subfamilies AluYe5, AluYe6 and AluYf5, the AluYe4, or the AluYe2, suggesting that chimeric sequences are frequent among Alus.

MATERIALS AND METHODS

Database of Alu *consensus* sequence

Alu *consensus* sequences were retrieved from databases and literature to construct the final collection of 87 sequences as follows: 47 from the Repbase Update (Jurka et al. 2005) and literature (Bennett et al. 2008; Park et al. 2005; Price et al. 2004; Styles and Brookfield 2007). The updated list of sequences is presented in Online Resource 1. In some cases, more than one *consensus* sequence is documented for the same subfamily (e.g. AluYa1_1 and AluYa1_2 correspond to two *consensus* sequences for the AluYa1 subfamily). To avoid arbitrary decisions, we included all the sequences in the database.

Sequence comparison and list of polymorphic sites

Alignment of the complete set of 87 Alu sequences was performed in Geneious v5.4 using the default options (Drummond et al. 2011). The AluJo *consensus* was set as reference sequence. Poly-A tails were removed from all sequences due to size heterogeneity. Sequence comparisons revealed a total of 146 polymorphic positions, of which, 12 are indels. The complete list of all polymorphic positions is provided in Online Resource 2. Position numbering was performed accordingly to AluJo (Fig. 1). Insertion and deletion polymorphisms (indels) are named

as in the following example: a single-base deletion in position 65 is indicated as “65delC” and an insertion of an adenine after position 177 is indicated as “177.1insA” as it represents a base insertion relative to the reference sequence (AluJo).

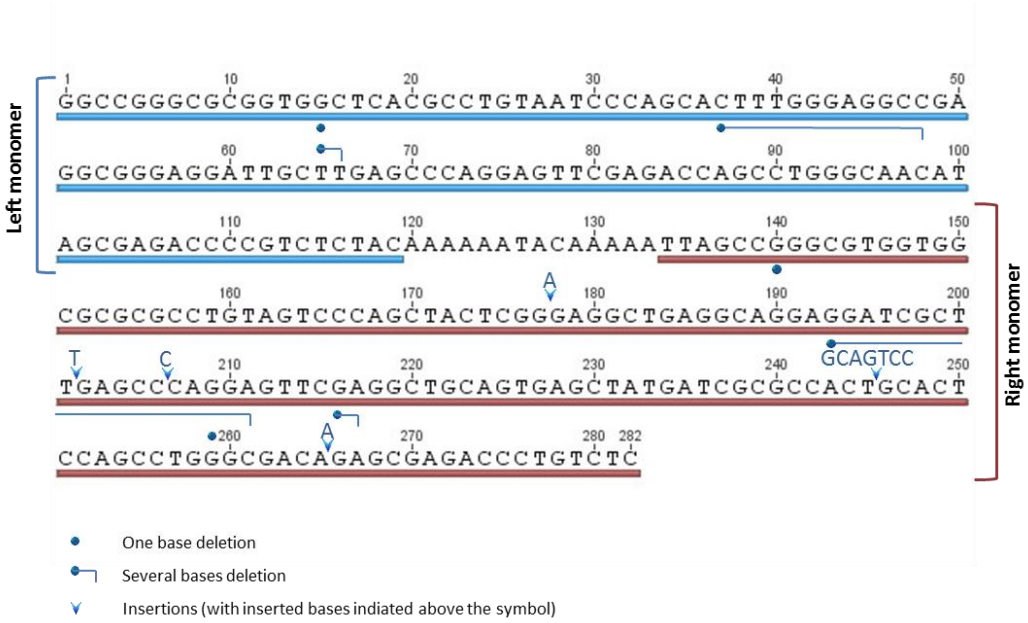


Fig. 1 Position of indel markers detected in the Alu *consensus* database relative to the AluJo *consensus* sequence (Jurka et al. 2005). The complete list of SNPs is provided in Online Resource 1.

Network construction

The Network 4610 software (http://www.fluxus-engineering.com/sharenet.htm) was used to construct the network based in all the 12 indels revealed by the comparison of the entire collection of Alu sequences. Allelic forma were converted in binary data (presence/absence) in the input file. The particular cases of positions 65delC and 65_66delCT were considered to be independently segregating sites. Poly-A linker and tail polymorphisms were not included. Each mutation site was equally weighted 10. The reduced median (RM) algorithm was tested with all the default parameters.

RESULTS

Database of polymorphic sites for *consensus* Alus

The collection of Alu *consensus* sequence retrieved from databases and related literature includes a total of 87 unique *consensus* sequences matching 74 distinct Alu subfamilies (Online Resource 1). Of these, four correspond to the ancestral AluJ, 20 are documented as AluS sequences and 50 as AluY, the youngest family member in primates (Mighell et al. 1997). Sequences were then aligned for further comparison after removing the poly-A tail, which would render the correct homology detection difficult, and compared with the reference (AluJo). A total of 146 polymorphic positions (SNPs and indels) were detected and combined into a single dataset

(Online Resource 2). This list of polymorphisms is expected to be useful for future research as it represents the most updated list of polymorphic sites of all known Alu *consensus* sequences. More than two alleles exist in most of the sites, strengthening that back and forward mutation are frequent events.

The polymorphic spectrum includes 12 indels with length sizes ranging from 1 to 19 bp (Fig. 1; Online Resource 2). With the exception of positions 65 and 66, there is no size heterogeneity, indicating they are useful markers to dissect the evolutionary history of Alu master genes.

The evolutionary history of human Alus

Taking advantage of indel markers found in the complete record of Alu *consensus* sequences in humans (Fig. 1; Online Resource 1) the network of haplotypic combinations was inferred as shown in fig. 2. With the exception of two reticulations (graphs identified as L and R in Fig. 2), that clearly demonstrate alternative solutions, the network is well resolved. The two reticulations observed (L and R) that link nodes 1, 2, 3, 4 and 7, 13, 14, 15, respectively, are unlikely to be the result of back mutation given the type of markers used in the network construction - indels. Instead, they might invoke events of recombination, a hypothesis that was further explored.

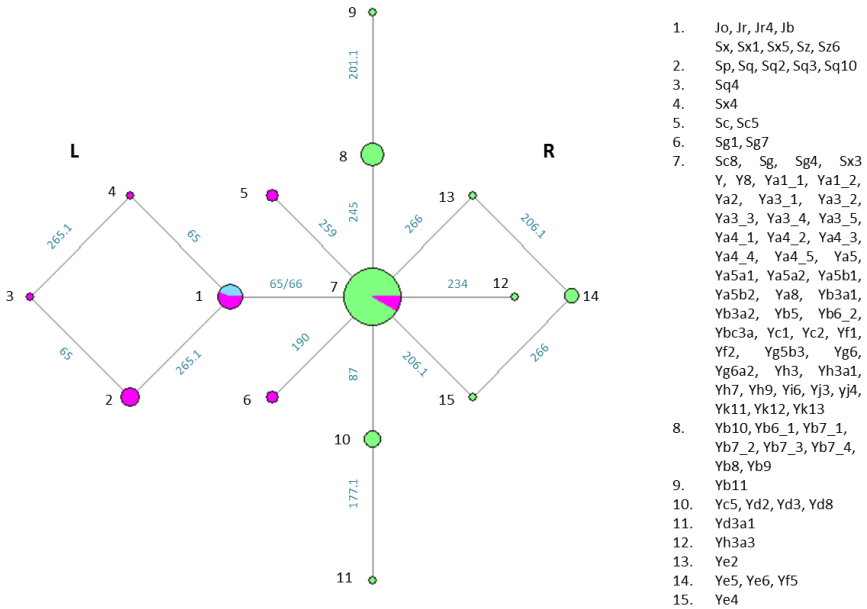


Fig. 2 Clustering of Alu subfamilies using indel (insertion/deletion) markers shown in Online Resource 2. The blue slice of node 1 represents the oldest subfamily (AluJ). AluS elements are represented in pink and members of the young AluY are shown in green. Indel sites are shown in branches. The two reticulations are indicated as L (left) and R (right).

In one of the cases (L), the Alu subfamilies represented in nodes 1, 2, 3 and 4 are distinguished by the haplotypic combination of 65/66 and 265.1 polymorphisms (Fig. 3). Four combinations were detected regarding the positions 65 and 66 (TT, CT, -T, --) located in the first monomer. Because positions 65 and 66 are deleted in the youngest AluY family when compared to the reference AluJo, the three remaining combinations (TT, CT, -T) are assumedly older. Hence, 65T/66T is the ancestral combination as it is observed in AluJ subfamily (Fig. 2, node 1) (Kapitonov and Jurka 1996). Following the same rationale, the 265.1insA at the second monomer was assumed to be the youngest allele. After the emergence of the 65C/66T combination, found in most AluS members, two alternative pathways are considered (Fig. 3, A and B) based on the order of mutational events occurring in each monomer.

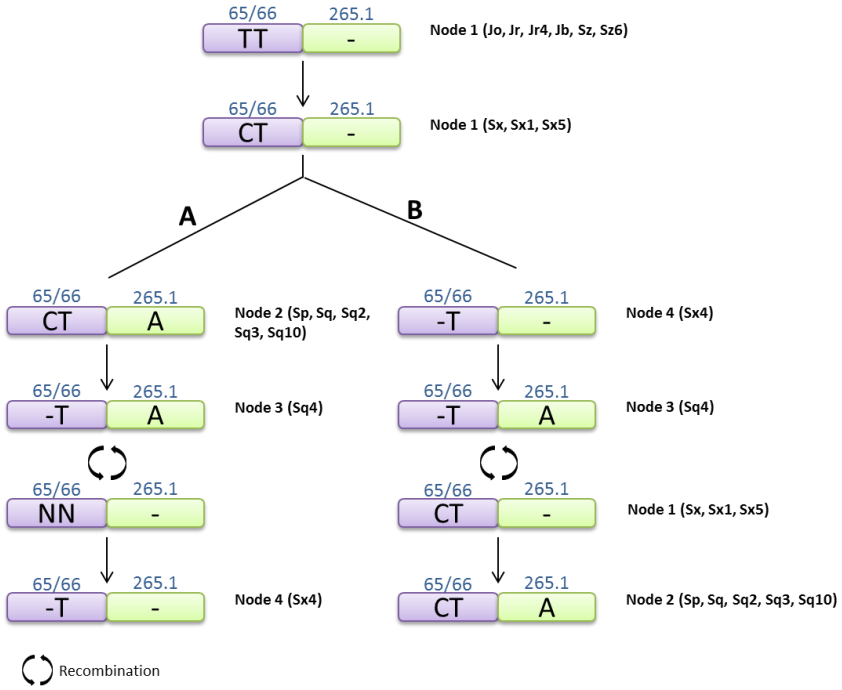


Fig. 3 Alternative pathways for the origin of Alu subfamilies clustered in nodes 2, 3, and 4 of Fig. 2. Left and right monomers are colored purple and green, respectively.

The first pathway (Fig. 3, A) illustrates the emergence of AluSp, AluSq, AluSq2, AluSq3 and AluSq10 (Fig. 3, node 2), AluSq4 (Fig. 3, node 3) and AluSx4 (Fig. 3, node 4) by an adenine insertion between 265 and 266 positions in any member of node 1 carrying the 65C/66T, thus originating Alus included in node 2. Then, the 65del in one of the Alus included in node 2 gave rise to the AluSq4 subfamily. Afterwards, a recombination event between the first monomer of AluSq4 and the second monomer of any Alu element (not carrying the 265.1insA) originated the novel AluSx4 subfamily. The alternative pathway (Fig. 3, B) assumes that the deletion in position 65 occurred before the 265.1insA. First, an element of node 1 fathered the AluSx4 subfamily by a

65delC, followed by the 265.1insA which generated the AluSq4 subfamily. Under this scenario, the subfamilies included in node 2 (e.g. AluSp) had origin in a recombination event between the right monomer of AluSq4 and the left monomer of any member of node 1 carrying the 65C/66T allele, that is to say, most of the AluS elements.

In-depth analyses of the sequences involved revealed that AluSx4 differs from the ancestral AluSq4 by the T98C substitution in the left monomer (Fig. 4). In addition, pairwise identity between the right monomer of all possible candidates to be donors, that is, those not carrying the 265.1insA, revealed that the most likely contributor was AluSx3 since both differ in a single site (G191A) (Fig. 4) and share 99.3% of sequence identity.

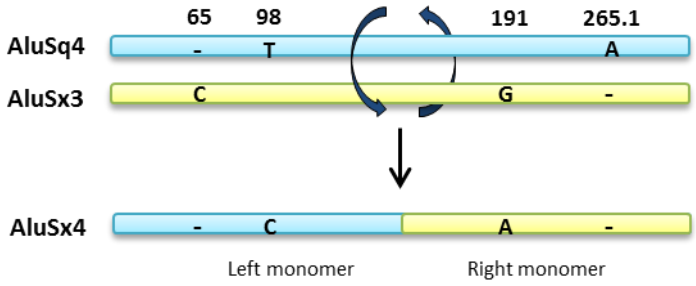


Fig. 4 Recombination event in the origin of AluSx4 master gene.

The second pathway (Fig. 3, B) is less likely as it would oblige a minimum of ten extra mutational steps subsequently to the putative recombination between AluSq4 and elements of node 1. Although both pathways involve a recombination event, the one that requires less mutational steps is the pathway A, which points to the origin of the AluSx4 subfamily throughout the recombination between an AluSq4 and any element carrying the 65C/66T allele (Fig. 3, fig. 4).

The second reticulation (Fig. 2, R) requires an even higher number of steps to be explained (Fig. 5). In this case, the key positions to establish the alternative mutational pathways followed after diverging from an ancestral Alu sequence are 206.1 and 266/267, both in the right monomer. These pathways are summarized as follows:

(A) Assuming that AluYe4 and AluYe2 resulted from distinct mutations (insertion of a C in 206.1 and deletion of a GA in position 266/267, respectively), of an ancestral sequence, and that a recombination event occurred between the first half of the right monomer of AluYe4 (node 15) and the second half of the right monomer of AluYe2 (node 13), members of node 14 (AluYe5, AluYe6 and AluYf5) represent an obligatory recombinant cluster.

(B) In this pathway, AluYe4 is a recombinant of the first half of the right monomer of AluYe5, AluYe6 or AluYf5 (node 14) and the second half of the right monomer of an ancestral Alu.

(C) AluYe2 (node 13) is a recombinant between the first half of the right monomer of an ancestral Alu and the second half of the right monomer of one of the AluYe5, AluYe6 or AluYf5 elements (node 14).

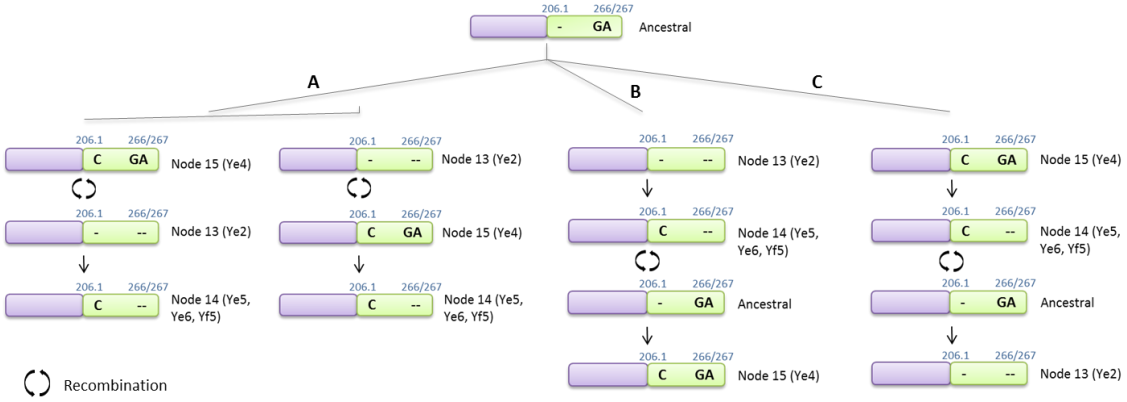


Fig. 5 Alternative pathways for the origin of Alu subfamilies clustered in nodes 13, 14 and 15 of Fig. 2. Left and right monomers are coloured purple and green, respectively. The ancestral sequence is any Alu with the indicated allelic combination in positions 206.1 and 266/267.

As with the previous example, the allelic configuration of these elements was analyzed and combined with information provided by pairwise identity scores between the involved elements. These analyses did not revealed the most parsimonious hypothesis, as the scores between recombinant (chimeric) Alus and their corresponding parental elements reached 100% or near 100% in all cases, which is the result of the recent origin of the AluY subfamily (Mighell et al. 1997). Notwithstanding, in all possible pathways described in Fig. 5, a recombination step is always required to explain the emergence of the observed haplotypes.

DISCUSSION

Alu elements are commonly found in primate genomes and it has been estimated that the average distance between any two Alus is approximately 3 Kb (Lander et al. 2001), although most of them are inactive, retrotransposition-competent elements. Events of ectopic recombination between Alu elements are known to be associated with deleterious rearrangements (Batzer and Deininger 2002; Conceicao Pereira et al. 2012; Deininger and Batzer 1999; Dobrovolny et al. 2011; Kuiper et al. 2011; Quental et al. 2009). Recombination is also known to create chimeric Alus (Johanning et al. 2003; Roy-Engel et al. 2002; Roy et al. 2000; Styles and Brookfield 2007) as are for instances those resurrected by partial gene conversion involving the poly-A tail at the 3'end (Johanning et al. 2003).

In this study, we searched for signals of recombination at the entire set of known Alu *consensus* sequences in order to broaden its effect in Alu evolution. To that, all known Alu *consensus* sequences were analyzed and compiled in a single file (Online Resource 1) that

includes 87 sequences from 74 subfamilies. A total of 146 polymorphisms were detected (Online Resource 2) and 12 indels used to establish the historical relationship between the distinct subfamilies. Two reticulations were observed in Fig. 2 that represents the graphical clustering of all 74 Alu subfamilies. After considering the possible pathways for the occurrence of nodes 2, 3 and 4 (Fig. 2, L) and nodes 13, 14 and 15 (Fig. 2, R) we could establish the role of recombination in the origin of the involved subfamilies. Our uncertainty in distinguishing between cross-over and gene conversion is due to the lack of information on the flanking genomic region of the original master genes. Although gene conversion has been assumedly more frequent that cross-over in Alu recombination (McVean 2010; Paigen and Petkov 2010), direct proof of gene conversion would only be possible if both recombination products are available (Chen et al. 2007).

A more general picture of the information provided by indels and SNPs allowed the distinction of Alu subfamilies according to stable positions (Fig. 6). Despite the information provided by the combination of both marker types, large clusters incorporating a vast number of subfamilies, mainly in what refers to young AluY elements, are still observed. It is important to mention that although a high number of segregating sites were detected among the Alu *consensus* sequences (Online Resource 2), only A120T, G194A, T214C, C215G and G219C represent single occurrences in the history of Alu *consensus* sequences (Fig. 6).

Data presented in Fig. 6 is relevant in many other aspects. The case of subfamily AluYc5 which shares with AluYd members the 87_98delATCCTGGCTAAC, and AluYf5 that shares with AluYe the 206.1insC allele, reveals that the boundaries of individualization of a subfamily are unclear. So, the questions we put forward are: (a) how many mutational steps should a source gene differ from its parental gene and still be considered as a subfamily member and, the other way around, (b) how many mutations are necessary to be considered as the founder of a new subfamily? Several subfamilies have been documented as having arisen from multiple *consensus* sequences (Matera et al. 1990), which further supports the need to consider multiple source genes in the birth of a novel subfamily. Furthermore, there is the need for a classification system that detects chimeric elements arising from ectopic recombination.

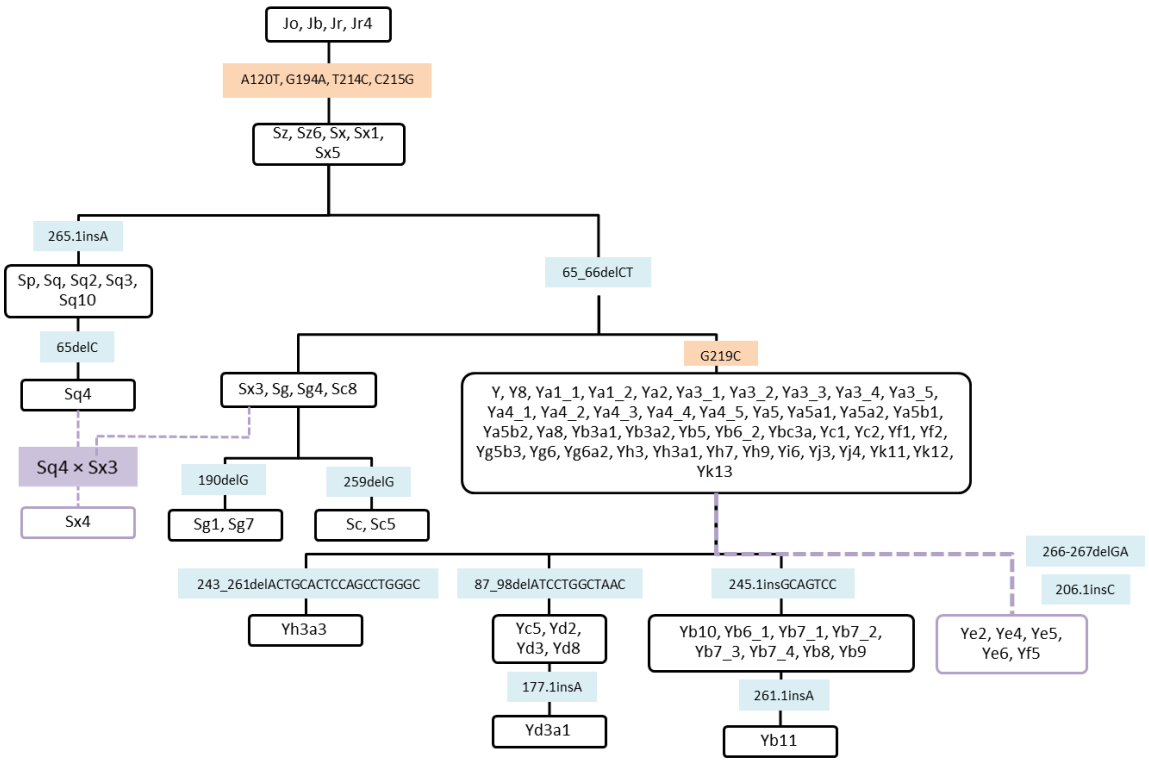


Fig. 6 Evolution of Alu subfamilies. Blue boxes are relative to indel (insertion/deletion) markers, orange boxes correspond to SNPs (single nucleotide polymorphisms) and purple boxes correspond to recombination events and recombinant subfamilies.

The expected fate of Alu elements is generalised in Fig. 7. An active Alu is retrotransposed without mutation (Fig. 7, A) or, with variants that do not impair retrotransposition ability (Fig. 7, B). In some other cases inactivating mutations impair the mobilization throughout the genome (Fig. 7, C). Recombination events between two Alus can result in truncated sequences (Fig. 7, D), but recombination can also occur between two active elements of differentiated subfamilies resulting in a chimeric, and still active Alu (Fig. 7, E). The two remaining cases (Fig. 7, F and G) are particularly important. We reinforce that recombination between an inactive and an active Alu can result in the birth of a novel active element (Fig. 7, F) and recombination between two inactive Alus that abolishes the inactivating mutation(s) may as well result in an active element. This points to the fact that ectopic recombination can resurrected inactive Alus allowing a new dimension in the dynamics that involves Alu inactivation and birth.

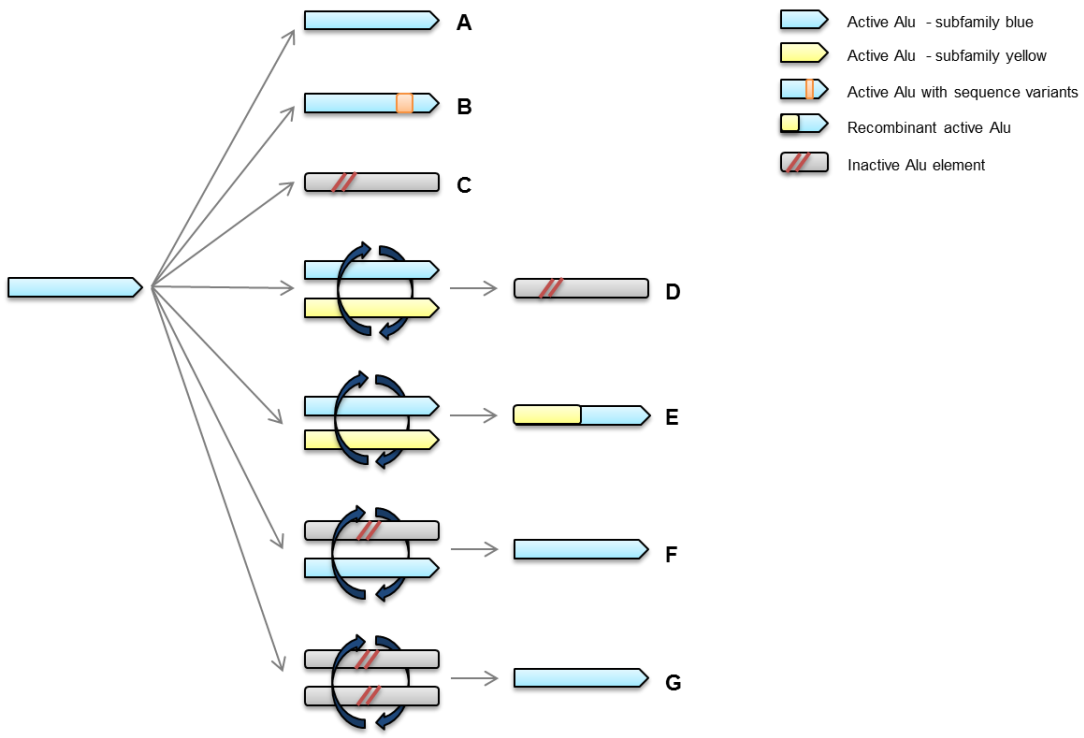


Fig. 7 Fate of Alu elements upon integration in the genome. (A) Alu is incorporated and remains active. (B) Alu inserted with mutations though remaining active. (C) Alu carrying inactivating mutations. (D) Recombination between active elements causing inactivation through the mutational process. (E) Recombination between two active elements and birth of a new, chimeric, subfamily. (F) Recombination between active and inactive elements and birth of an active Alu. (G) Possible resurrection of two inactive Alus by recombination.

CONCLUSIONS

The motivation behind the work here presented is the search for a better understanding of the role of recombination during the evolution of Alu sequences that punctuate the human genome. In this sense, the importance of designing a robust classification system to detect recombinant Alu elements is underlined here. Alu categorization based only on the traditional diagnostic mutation is insufficient as many of the so-called diagnostic positions were shown to be shared by distinct subfamily members. An alternative, which was here demonstrated to be useful, is the combination of point mutations with more stable markers such as indels. Finally, our results show that the role of Alu ectopic recombination in the origin of novel, chimeric, Alu subfamilies is expected to have vast implications in the clinical, evolutionary and forensic fields.

ACKNOWLEDGEMENTS

LA and RMS are supported by FCT through the program Ciencia2007 (Hiring of PhDs for the SCTN - financed by POPH - QREN - Tipology 4.2 - Promoting Scientific Employment, co-financed by MCTES national funding and The European Social Fund) and the research project PTDC/BIA-PRO/099888/2008.

REFERENCES

Batzer MA, Deininger PL (2002) Alu repeats and human genomic diversity. Nature Reviews Genetics 3: 370-379

Bennett EA, Keller H, Mills RE, Schmidt S, Moran JV, Weichenrieder O, Devine SE (2008) Active Alu retrotransposons in the human genome. Genome Research 18: 1875-1883

Chen J-M, Cooper DN, Chuzhanova N, Ferec C, Patrinos GP (2007) Gene conversion: mechanisms, evolution and human disease. Nature Reviews Genetics 8: 762-775

Comeaux MS, Roy-Engel AM, Hedges DJ, Deininger PL (2009) Diverse cis factors controlling Alu retrotransposition: What causes Alu elements to die? Genome Research 19: 545-555

Conceicao Pereira M, Loureiro JL, Pinto-Basto J, Brandao E, Margarida Lopes A, Neves G, Dias P, Geraldes R, Martins IP, Cruz VT, Kamsteeg E-J, Brunner HG, Coutinho P, Sequeiros J, Alonso I (2012) Alu elements mediate large SPG11 gene rearrangements: further spatacsin mutations. Genetics in medicine : official journal of the American College of Medical Genetics 14: 143-51

Deininger PL, Batzer MA (1999) Alu repeats and human disease. Molecular Genetics and Metabolism 67: 183-193

Dobrovolny R, Nazarenko I, Kim J, Doheny D, Desnick RJ (2011) Detection of Large Gene Rearrangements in X-linked Genes by Dosage Analysis: Identification of Novel alpha-Galactosidase A (GLA) Deletions Causing Fabry Disease. Human Mutation 32: 688-695

Drummond A, Ashton B, Buxton S, Cheung M, Cooper A, Duran C, Field M, Heled J, Kearse M, Markowitz S, Moir R, Stones-Havas S, Sturrock S, Thierer T, Wilson A (2011) Geneious v5.4, available from http://www.geneious.com/

Han K, Lee J, Meyer TJ, Wang J, Sen SK, Srikanta D, Liang P, Batzer MA (2007) Alu recombination-mediated structural deletions in the chimpanzee genome. Plos Genetics 3: 1939-1949

Johanning K, Stevenson CA, Oyeniran OO, Gozal YM, Roy-Engel AM, Jurka J, Deininger PL (2003) Potential for retroposition by old Alu subfamilies. Journal of Molecular Evolution 56: 658-664

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J (2005) Repbase update, a database of eukaryotic repetitive elements. Cytogenetic and Genome Research 110: 462-467

Kapitonov V, Jurka J (1996) The age of Alu subfamilies. Journal of Molecular Evolution 42: 59-65

Kuiper RP, Vissers LELM, Venkatachalam R, Bodmer D, Hoenselaar E, Goossens M, Haufe A, Kamping E, Niessen RC, Hogervorst FBL, Gille JJP, Redeker B, Tops CMJ, van Gijn ME, van den Ouweland AMW, Rahner N, Steinke V, Kahl P, Holinski-Feder E, Morak M, Kloor M, Stemmler S, Betz B, Hutter P, Bunyan DJ, Syngal S, Culver JO, Graham T, Chan TL, Nagtegaal ID, van Krieken JHJM, Schackert HK, Hoogerbrugge N, van Kessel AG, Ligtenberg MJL (2011) Recurrence and Variability of Germline EPCAM Deletions in Lynch Syndrome. Human Mutation 32: 407-414

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860-921

Matera AG, Hellmann U, Hintz MF, Schmid CW (1990) Recently transposed Alu repeats result from multiple source genes

Nucleic Acids Research 18: 6019-6023

McVean G (2010) What drives recombination hotspots to repeat DNA in humans? Philosophical Transactions of the Royal Society B-Biological Sciences 365: 1213-1218

Mighell AJ, Markham AF, Robinson PA (1997) Alu sequences. Febs Letters 417: 1-5

Myers JS, Vincent BJ, Udall H, Watkins WS, Morrish TA, Kilroy GE, Swergold GD, Henke J, Henke L, Moran JV, Jorde LB, Batzer MA (2002) A comprehensive analysis of recently integrated human Ta L1 elements. American Journal of Human Genetics 71: 312-326

Paigen K, Petkov P (2010) Mammalian recombination hot spots: properties, control and evolution. Nat Rev Genet 11: 221-33

Park ES, Huh JW, Kim TH, Kwak KD, Kim W, Kim HS (2005) Analysis of newly identified low copy AluYj subfamily. Genes & Genetic Systems 80: 415-422

Price AL, Eskin E, Pevzner PA (2004) Whole-genome analysis of Alu repeat elements reveals complex evolutionary history. Genome Research 14: 2245-2252

Quental R, Azevedo L, Rubio V, Diogo L, Amorim A (2009) Molecular mechanisms underlying large genomic deletions in ornithine transcarbamylase (OTC) gene. Clinical Genetics 75: 457-464

Rogers JH (1985) The origin and evolution of retroposons. International Review of Cytology-a Survey of Cell Biology 93: 187-279

Roy-Engel AM, Carroll ML, El-Sawy M, Salem AH, Garber RK, Nguyen SV, Deininger PL, Batzer MA (2002) Non-traditional Alu evolution and primate genomic diversity. Journal of Molecular Biology 316: 1033-1040

Roy AM, Carroll ML, Nguyen SV, Salem AH, Oldridge M, Wilkie AOM, Batzer MA, Deininger PL (2000) Potential gene conversion and source genes for recently integrated Alu elements. Genome Research 10: 1485-1495

Stoneking M, Fontius JJ, Clifford SL, Soodyall H, Arcot SS, Saha N, Jenkins T, Tahir MA, Deininger PL, Batzer MA (1997) Alu insertion polymorphisms and human evolution: Evidence for a larger population size in Africa. Genome Research 7: 1061-1071

Styles P, Brookfield JFY (2007) Analysis of the features and source gene composition of the AluYg6 subfamily of human retrotransposons. Bmc Evolutionary Biology 7

Ullu E, Tschudi C (1984) Alu sequences are processed 7SL RNA genes. Nature 312: 171-172

Weiner AM, Deininger PL, Efstratiadis A (1986) Nonviral retroposons-genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information Annual Review of Biochemistry 55: 631-661

Willard C, Nguyen HT, Schmid CW (1987) Existence of at least three distinct Alu subfamilies Journal of Molecular Evolution 26: 180-186

Zhi D (2007) Sequence correlation between neighboring Alu instances suggests post-retrotransposition sequence exchange due to Alu gene conversion. Gene 390: 117-121

SUPPLEMENTARY FILES (provided in electronic format)

Online Resource 1

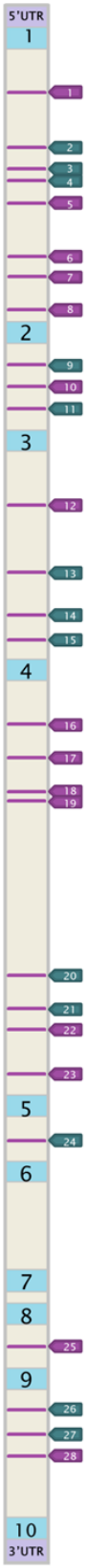
Updated list of Alu *consensus* sequences as retrieved from Repbase Update (Jurka et al. 2005) and the literature (Bennett et al. 2008; Park et al. 2005; Price et al. 2004; Styles and Brookfield 2007).

Online Resource 2

The complete list of all polymorphic positions detected in the complete list of Alu *consensus* sequences. Position numbering was performed accordingly to AluJo (Fig. 1). Major subfamily-specific mutations represented in Fig. 6 are colored blue (sites 120, 194, 214 and 215) and green (site 219) and are specific of AluJ and AluY, respectively.

SECTION II

The *OTC* Alus



A whole gene scan of *OTC* gene revealed a total of 28 Alu elements (Figure 13). Their sequences were aligned to allow a better visualisation of their structure (Figure 14). The distribution of these Alu elements between the sense and the antisense strands is similar and widespread. The pairwise identity between any two Alus ranges from 72.4% to 89.2% (Table 3, Table 4 and Table 5Table 5) in accordance to the general expected values observed in the primate genome. About 12% of the OTC sequence is occupied by Alu sequences, that is to say, one Alu insertion every 2.5 Kb, once again in accordance with the expected values of the Alu density for the whole genome (one insertion per 3 Kb).

Two deletions in the *OTC* gene involving two pairs of Alu elements mapped in this work were previously described [137]. The first case was a deletion of exon 2, involving Alu 6 and Alu 11 inserted in opposite directions, probably by the mechanism described above in Figure 7; the second reported case was a deletion of exons 6 to 9, concerning Alu 24 and Alu 26 both inserted in the anti-sense strand, by the mechanism illustrated in Figure 8.

Since the distribution of Alus inserted forward and reversely is balanced within the gene in a proportion of 15 Alus forward to 13 Alus reverse, and the values of pairwise identity between Alus in the same (Table 3 and Table 4) or in opposite strands (Table 5) are very similar, all the recombination mechanisms of recombination Alu-mediated are conceivable, possibly leading to rearrangements that may or may not have phenotypic consequences.

Figure 13: Relative location of the 28 Alus within the intronic regions of the *OTC* gene. Light blue boxes represent the 10 exons; pink and green tags refer to forward and reversely inserted Alus.

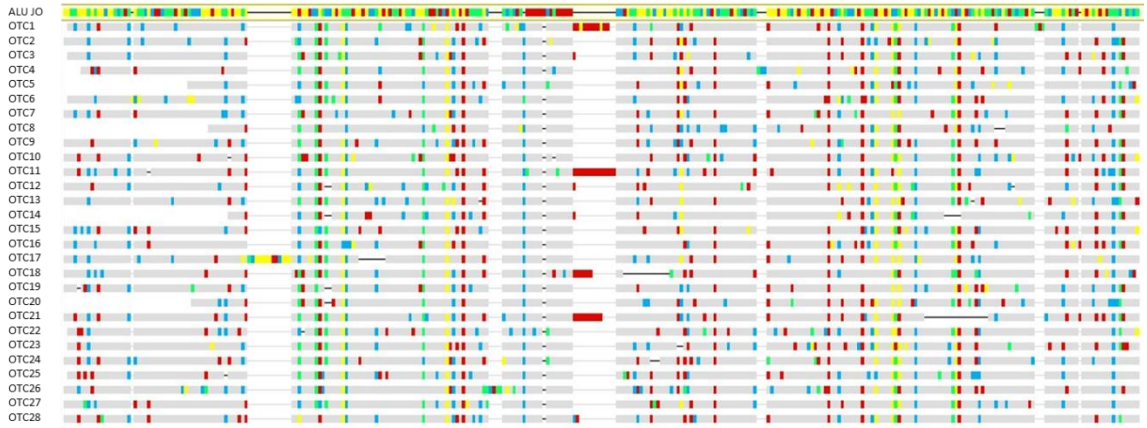


Figure 14: *OTC* Alus alignment using the *consensus* AluJo as reference.

Table 3: Percentage of pairwise identity between any two Alus inserted in the sense strand.

	Alu2	Alu3	Alu4	Alu9	Alu11	Alu13	Alu14	Alu15	Alu20	Alu21	Alu24	Alu26
Alu3	84,3											
Alu4	80,9	82,7										
Alu9	80	83,7	81,5									
Alu11	79,5	82,7	77,2	81,8								
Alu13	81,9	84,4	81	82,4	79,9							
Alu14	80,7	81,4	80,9	82,1	80,3	78,6						
Alu15	82	83,7	81,8	80,9	81,5	79,7	81,3					
Alu20	79,6	82,2	80,5	80,5	78,5	80,6	80,7	81,3				
Alu21	78,6	81,6	77,3	76,8	80,6	76,3	77,8	78,8	75			
Alu24	79	80,6	81,5	81,3	77,9	81,1	78,1	82,2	79,8	75,2		
Alu26	79,3	81,9	80,8	82,2	77,2	78,7	78,4	79,6	76,3	73,9	80,2	
Alu27	82,1	83,8	82,5	80,9	80,6	81,6	79,9	82,8	80	75,9	81,4	78,8

Table 4: Percentage of pairwise identity between any two Alus inserted in the anti-sense strand.

	Alu1	Alu5	Alu6	Alu7	Alu8	Alu10	Alu12	Alu16	Alu17	Alu18	Alu19	Alu22	Alu23	Alu25
Alu5	79,2													
Alu6	78,2	80,6												
Alu7	79,9	80,8	79,4											
Alu8	77,8	81,6	77	80,2										
Alu10	76,5	78,1	79	80,2	80,6									
Alu12	78,1	80,1	77,9	81,8	78,4	81,1								
Alu16	80,1	82,1	82,1	84,5	80,2	84,6	82,9							
Alu17	76,9	77,2	74,2	77,1	73,5	73,1	76,5	77,7						
Alu18	77,3	76,2	75,3	77,9	75,4	78,6	78,3	78	72,8					
Alu19	79,4	80,9	80,2	86,1	80,4	82,7	82,5	83,6	77,2	76,3				
Alu22	76,2	80,8	80	80,9	79	81,2	79,6	82	74,9	78,5	81,2			
Alu23	78	77,9	78,1	81,6	75,7	78,8	78,1	80,4	76,4	75,6	82,1	81,6		
Alu25	79,2	80,7	81	80,8	79,8	80,7	80,3	82,9	76,3	78,6	82,9	82	78,2	
Alu28	80,1	81,6	78,7	80,9	78,1	82,2	80,8	83,6	75,1	80,3	81,1	83,8	79,5	83

Table 5: Percentage of pairwise identity between any two Alus inserted in opposite strands.

	Alu1	Alu5	Alu6	Alu7	Alu8	Alu10	Alu12	Alu16	Alu17	Alu18	Alu19	Alu22	Alu23	Alu25	Alu28
Alu2	80,6	81,9	80	79,9	77,5	80,2	81	83,3	77,6	76,8	82,4	80,1	79,2	81,4	81,5
Alu3	80,3	82,9	81,5	83,8	80,8	86,7	82,7	85,9	78	78,9	84,2	83,1	81,1	82,7	83,1
Alu4	80,1	82	83,1	82,7	80,9	81,2	80,7	83,5	77	77,2	82,5	79,3	78,7	81,9	80,5
Alu9	78,8	81,7	78,2	84,1	79,3	80	81,9	81,9	75,8	79,3	82,2	80,9	81,7	80,2	82,1
Alu11	81,9	79,3	76	80,2	77,4	80,5	81,2	82,2	75,4	80,3	79,9	79,7	79,4	77,3	82,5
Alu13	77,9	80,8	79,4	81,1	77,9	81,7	81,5	83,4	79,4	77	84,5	79,9	80,6	79,7	81,6
Alu14	77,3	81,3	78,6	79,8	78,9	76,9	89,2	80,2	74,3	75,2	82,3	78,5	76,9	78,6	80,7
Alu15	81,6	83,6	79,8	83,4	80,5	80,7	81,9	83,5	78,6	78,6	81,5	82,3	81,6	79,9	83,3
Alu20	77,5	79,3	79,1	82,2	81,3	79	82	79,1	75	75,5	84,1	80,2	76,3	79,8	80
Alu21	78	76,6	73,2	79,4	74,5	77	77,5	79,4	72,4	75,5	78,1	74,6	73	74,5	77,2
Alu24	79,4	82,4	77,6	82,8	79,7	80,3	79	79	77,4	78	84,2	79,9	82,2	80,2	79,8
Alu26	77,6	79,7	80,9	82,7	79,7	79,6	80	81	73,9	76,2	80,5	80,5	79,9	82,7	79,4
Alu27	80,6	80,6	78,3	80,6	78,9	80,2	80,6	81,3	77,6	80,3	81,1	80,7	78,5	81,3	83,8

Table 6: Resulting classification provided by different software tools (Repeat Masker, CENSOR and CAlu) for the 28 Alus of the human *OTC* gene. Indel-based network correspond to the classification system developed in this project as indicated in the section I of the results.

Alu	Repeat Masker	CENSOR	CAlu	Indel-based Network
1	AluSx	AluSx	AluSx	Unclassified
2	AluSx	AluSx1	AluSp	Node 1
3	AluSx	AluSq2	AluSp	Node 2
4	AluSx	AluSq2	AluSq	Node 2
5	AluSx	AluSq	AluSx	Node 1
6	AluSx	AluSx	AluSx	Node 1
7	AluSx	AluSz	AluSz	Node 1
8	AluSx	AluSz	AluSz	Node 1
9	AluSx	AluSz	AluSz	Node 1
10	AluSx	AluSp	AluSp	Node 2
11	AluSx	AluS	AluSx	Node 1
12	AluY	AluY	AluY	Node 7
13	AluSx	AluSx1	AluSx	Node 1
14	AluY	AluY	AluY	Node 7
15	AluSx	AluSx1	AluSx	Node 1
16	AluSx	AluSp	AluSp	Node 2
17	AluSx	AluS	AluSz	Node 1
18	AluSx	AluSx	AluSx	Node 1
19	AluSx	AluSq	AluSg	Node 7
20	AluSx	AluSq	AluSg	Node 7
21	AluSx	AluSq2	AluSq	Node 2
22	AluSx	AluSx	AluSx	Node 1
23	AluSx	AluSx	AluSx	Node 1
24	AluSx	AluSx	AluSx	Node 1
25	AluSx	AluSz	AluSz	Node 1
26	AluSx	AluSz	AluSz	Node 1
27	AluSx	AluSx1	AluSx	Node 1
28	AluSx	AluSz	AluSz	Node 1

The highest value of pairwise identity between all Alus found in the *OTC* gene is relative to the pair Alu 12 and Alu 14, indicating that these Alus are evolutionary recent and belong to the Alu Y subfamily. In line with this, an attempt to classify *OTC* Alus was made, using the current

classification system and the widely used softwares Repeat Masker, CENSOR and CAlu. The results are shown in Table 6.

It is noticeable that for most cases (Alus 2, 3, 4, 5, 7, 8, 9, 10, 16, 17, 19, 20, 21, 25, 26 and 28) the results are not concordant, and in some cases, results are even conflicting in all three softwares (Alu 19 and Alu 20).

As stated before, current systems of classification are mainly based on SNPs, which often undergo recurrence within these elements, or even in polymorphisms located in the poly-A middle linker, a rather unstable region, leading to the misclassification of Alus. For this reason, a network containing not only all the known *consensus* sequences' haplotypes, but also those from *OTC* Alus was built (Figure 15), based on the same 12 indel markers mentioned in the manuscript "The role of recombination in the emergence of novel Alu subfamilies" presented above.

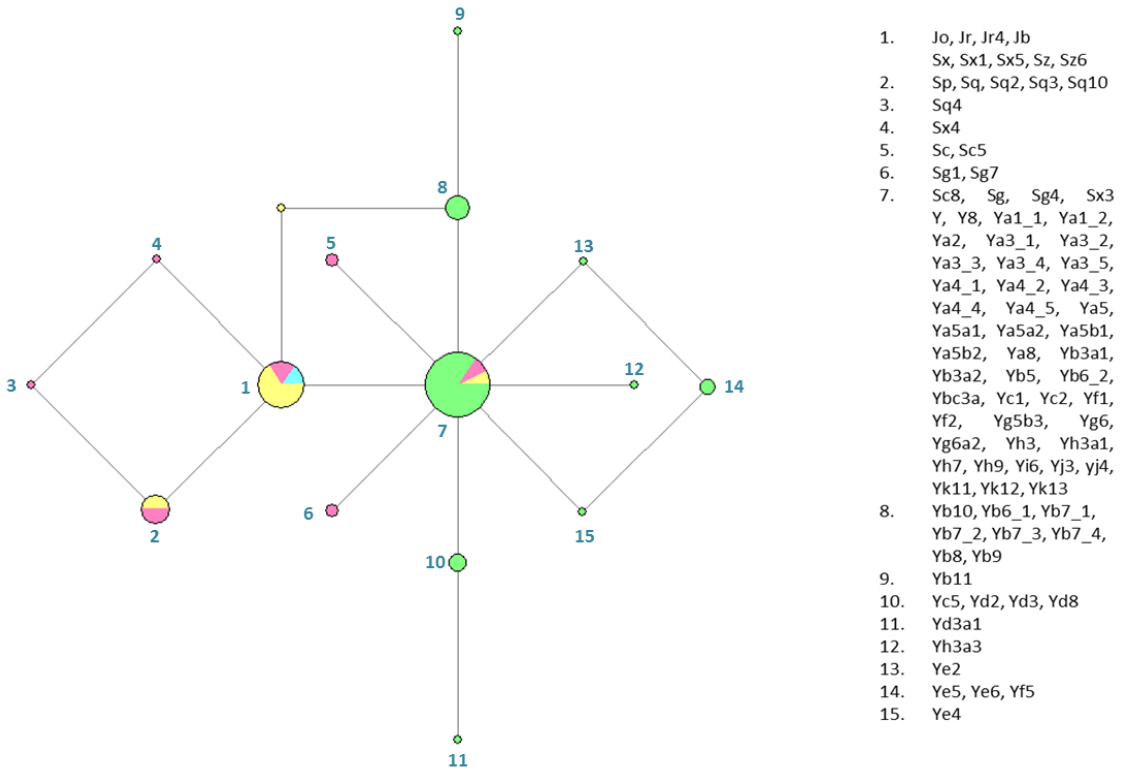


Figure 15: Network of all known Alu *consensus* sequences and *OTC* Alus. The blue slice represents AluJ. Pink, green and yellow slices and nodes represent AluS, AluY and the *OTC* Alus, respectively.

Herein, *OTC* Alus were clustered into 4 categories, corresponding to three subfamilies clusters and another separate cluster whose indel haplotype does not correspond to any documented subfamily. *OTC* Alus clusters are shown in Table 6. The un-clustered *OTC* Alu (Alu 1)

is likely the result of a recombination between an element from node 1 with another from node 8, since it shares diagnostic indel markers from both of these clusters. This is unlikely the result of back mutation, given the stable nature of these markers (Figure 16).

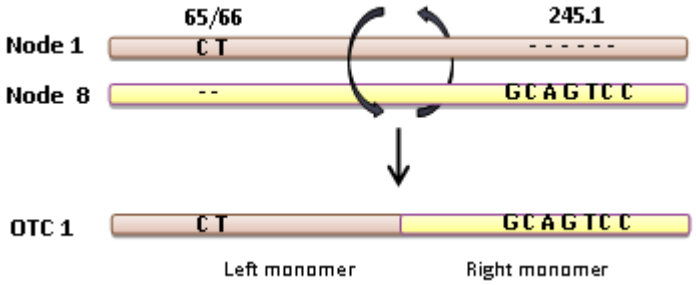


Figure 16: Possible recombination event behind the origin of the Alu *OTC* 1.

An accurate Alu classification system can be very useful for easily determining the age of an element, its origin, and may even help to unveil taxonomic questions. From the results stated above, one can conclude that the current classification system is very flawed. This system is not only based on hyper-mutable diagnostic positions, but also does not foresee recombination events between Alu sequences. Furthermore, there are too many lower levels of subfamilies that, in practical terms, do not have any utility whatsoever. A robust system of classification should be based on more stable markers such as indels or stable SNPs, and the classification of each Alu cluster should be done according to the elements' evolutionary path, instead of the order of publication. Moreover, the detection of recombinant Alus would constitute a major advance in the classification system, since recombination may be associated with important genomic rearrangements.

On the other hand, any classification system is dependent on a correct alignment, especially when indels are involved. Additionally, a classification system based on indels would only have two main limitations: the difficulty in narrowing more recent evolutionary events, due to their low mutation rate, and the impossibility to define indel markers close to the Alu extremity because of the frequent 5' and 3' Alu truncations. So, the ideal system would have to be based on a combination of markers with different mutation rates and located in regions not affected by end truncation.

OTC indel haplotypes

A successful multiplex-based strategy was designed for the simultaneously amplification of six OTC markers (Figure 17). This is important because it will assist current methodologies in the detection of OTC rearrangements. The allelic frequency of each polymorphic site is shown in

Table 1. The frequency of phased alleles (haplotypes) is shown in Table 7 and Figure 18. As observed, the two most frequent haplotypes, H3 and H7, account for 60% of haplotypes, yet none of them carry the combination of the most frequent allele of each marker.

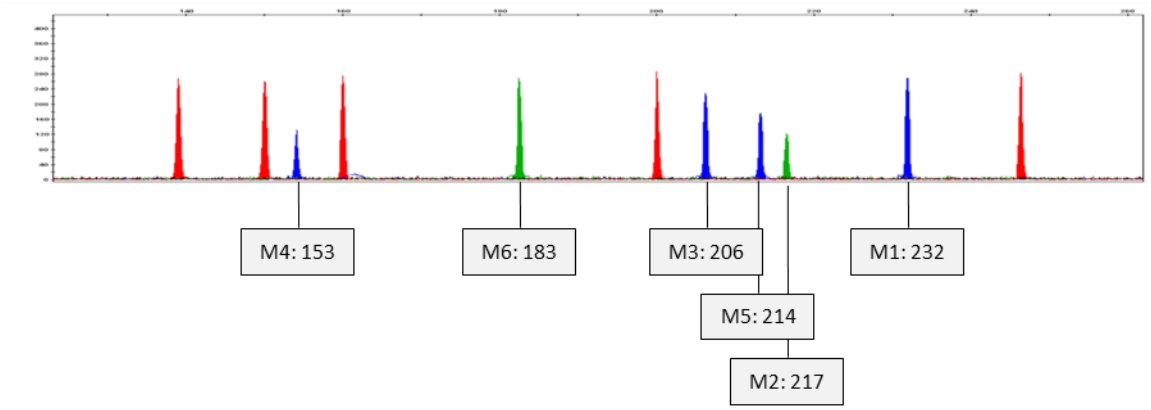


Figure 17: Example of a male profile obtained by capillary electrophoresis of the multiplex-system based in six *OTC* intronic markers (blue and green labeled). Molecular marker is labeled red (ROX 500).

Table 7: Haplotypes frequencies in the European Caucasian Population.

Haplotype	M1	M2	M3	M4	M5	M6	Frequency (N=38)
H1	232	211	200	154	213	185	0.13
H2	232	211	206	154	213	185	0.03
H3	232	217	206	153	214	183	0.3
H4	232	217	206	154	214	183	0.03
H5	232	217	206	154	213	185	0.13
H6	236	217	206	153	214	183	0.11
H7	236	217	206	154	213	185	0.3

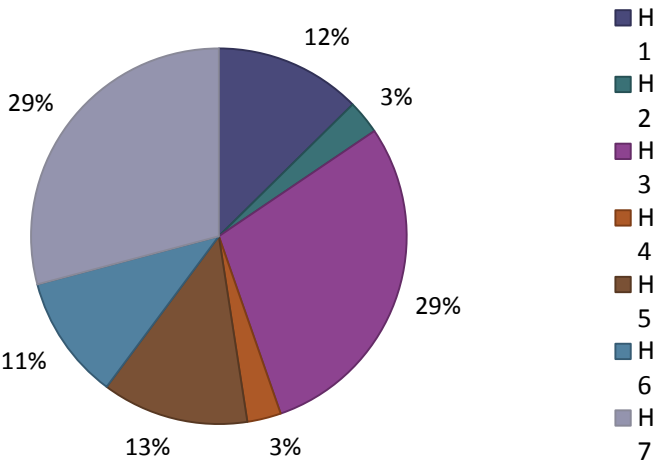


Figure 18: Haplotypic frequencies in the European Caucasian population.

OTC recombination hotspot

Patterns of linkage disequilibrium between the markers and the number of observed haplotypes (7) revealed a recombination hotspot within the gene (Figure 19), between the markers M3 and M4. Studies will be performed in order to understand the sequence-context of the recombination event. In this case, the recombination mechanism behind this dissociation is a crossover, whose breakpoint is located somewhere between intron 4 and intron 6.

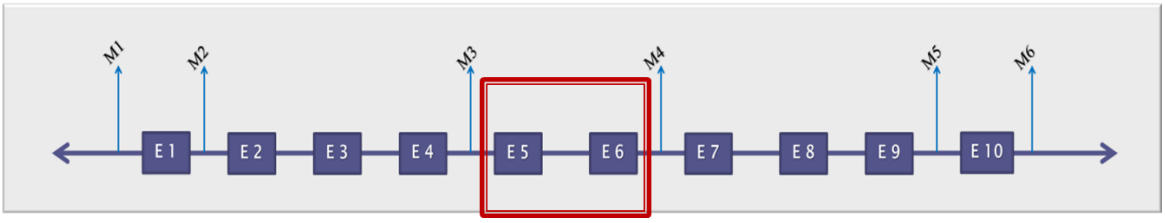


Figure 19: Relative position of the crossover point within the *OTC* gene (red box).

It is interesting that this recombination hotspot corresponds to the areas of the *OTC* gene with the lowest Alu density, indicating that recombination in this location may not be Alu-mediated, or may have resulted in Alu depletion. Furthermore, this point is not located near the recombinant Alu OTC1, indicating that OTC1's chimerisation was the result of gene conversion or a crossover that did not result in linkage dissociation. In-dept sequence analyses will need to be performed in order to detect the exact DNA sequence involved, considering this region of the gene is a hotspot for point mutations. Because most *OTC* mutations are *de novo* (i.e. germinal), it would be interesting to assess the role of recombination in this mutational mechanism.

Conclusions and Future Perspectives

During the course of this project a number of topics were addressed. In the matter of Alu-mediated recombination, we were able to broaden the knowledge of the evolutionary pathway of Alu subfamilies, detecting two subfamilies that arose from recombination events and further clarifying the importance of these events in the clinical, evolutionary and forensic fields. Altogether, the results reinforced the defectiveness of the Alu classification systems and current available bioinformatic tools, and special emphasis was given to the importance of designing a robust classification system based on stable markers that allow recombinant Alus detection. These results have already been submitted for publication, and the manuscript is incorporated in the chapter "Results and Discussion".

A genomic region that is known to undergo Alu-mediated rearrangements is the *OTC* gene. In an attempt to locate and classify *OTC* Alus, the whole gene was scanned. These analyses revealed an *a priori* probability of Alu-mediated rearrangements equally distributed throughout the gene. In addition, neutral polymorphic indel markers in the non-coding region of the *OTC* gene were identified and a multiplex-based system involving these markers was designed for the purpose of detecting large rearrangements in OTCD patients. The new polymorphisms showed to be distributed along the entire genomic region and can thus be used to define the haplotypic structure of the normal population and whenever such association is disrupted in a disease-associated rearrangement. Moreover, the study of these phased alleles also assisted in locating recombination points within the gene.

With all these questions answered, many other arose, encouraging additional research in this topic. The foremost important line of research would be the reclassification of the Alu elements, taking into account stable diagnostic positions, and the development of databases and programs for this new classification system to allow chimeric Alus detection.

Regarding additional forensic applications, as gene conversion events are known to be responsible for subfamily changes within the same *locus*, Alu elements could be used as subfamily polymorphisms, in addition to the insertion polymorphisms. Furthermore, species-specific subfamilies can be used to narrow the classification of primate taxa, or even to discriminate human from non-human samples based on human-specific subfamilies. Since Alus are in close proximity with coding regions, another interesting line of investigation would be trying to establish a relationship between the currently used set of Alu insertion polymorphisms in human individual identification with phenotypic characteristics associated with the Alu linked genes, in order to avoid the unethical use of these markers.

Regarding the study of the *OTC* Alus, further studies can be made, for instance, to widen the analyses of Alus in the entire region of the X chromosome, with a special focus on the rearrangement hotspot region Xp21.2-11.4. Other interesting perspectives would be the study of the retrotransposition ability and the evaluation of the polymorphic status of the *OTC* Alus.

References

1. Britten and Kohne (1968) Repeated sequences in DNA. *Science*, 161:529-&.
2. Mcclintock (1956) Controlling elements and the gene. *Cold Spring Harbor Symposia on Quantitative Biology*, 21:197-216.
3. Craig, Craigie, Gellert and Lambowitz, *Mobille DNA II.* Washington DC: ASM Press; 2002.
4. Lander, Linton, Birren, Nusbaum, Zody*, et al* (2001) Initial sequencing and analysis of the human genome. *Nature*, 409:860-921.
5. Beck, Collier, Macfarlane, Malig, Kidd*, et al* (2010) LINE-1 retrotransposition activity in human genomes. *Cell*, 141:1159-U1110.
6. Dewannieux, Esnault and Heidmann (2003) LINE-mediated retrotransposition of marked Alu sequences. *Nature Genetics*, 35:41-48.
7. Mills, Bennett, Iskow and Devine (2007) Which transposable elements are active in the human genome? *Trends in Genetics*, 23:183-191.
8. Belancio, Hedges and Deininger (2008) Mammalian non-LTR retrotransposons: For better or worse, in sickness and in health. *Genome Research*, 18:343-358.
9. Callinan and Batzer (2006) Retrotransposable elements and human disease. *Genome and Disease*, 1:104-115.
10. Swergold (1990) IDENTIFICATION, CHARACTERIZATION, AND CELL SPECIFICITY OF A HUMAN LINE-1 PROMOTER. *Molecular and Cellular Biology*, 10:6718-6729.
11. Babushok and Kazazian (2007) Progress in understanding the biology of the human mutagen LINE-1. *Human Mutation*, 28:527-539.
12. Szak, Pickeral, Makalowski, Boguski, Landsman*, et al* (2002) Molecular archeology of L1 insertions in the human genome. *Genome biology*, 3:research0052.
13. Ostertag, Goodier, Zhang and Kazazian (2003) SVA elements are nonautonomous retrotransposons that cause disease in humans. *American Journal of Human Genetics*, 73:1444-1451.
14. Wang, Xing, Grover, Hedges, Han*, et al* (2005) SVA elements: A hominid-specific retroposon family. *Journal of Molecular Biology*, 354:994-1007.
15. Deininger and Daniels (1986) The recent evolution of mammalian repetitive DNA elements. *Trends in Genetics*, 2:76-80.
16. Ullu and Tschudi (1984) Alu sequences are processed 7SL RNA genes. *Nature*, 312:171-172.
17. Quentin (1992) Fusion of a free left Alu monomer and a free right Alu monomer at the origin of the Alu family in the primate genomes. *Nucleic Acids Research*, 20:487-493.
18. Quentin (1992) Origin of the Alu family - a family of Alu-Like monomers gave birth to the left and the right arms of the Alu elements. *Nucleic Acids Research*, 20:3397-3401.
19. Batzer and Deininger (2002) Alu repeats and human genomic diversity. *Nature Reviews Genetics*, 3:370-379.
20. Fuhrman, Deininger, Laporte, Friedmann and Geiduschek (1981) Analysis of transcription of the human Alu family ubuquitous repeating element by eukaryotic RNA polymerase-III. *Nucleic Acids Research*, 9:6439-6456.
21. Knight, Batzer, Stoneking, Tiwari, Scheer*, et al* (1996) DNA sequences of Alu elements indicate a recent replacement of the human autosomal genetic complement. *Proceedings of the National Academy of Sciences of the United States of America*, 93:4360-4364.
22. Novick, Batzer, Deininger and Herrera (1996) The mobile genetic element Alu in the human genome. *Bioscience*, 46:32-41.
23. Perezstable, Ayres and Shen (1984) Distinctive sequence organization and functional programming of an Alu repeat promoter. *Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences*, 81:5291-5295.
24. Arcot, Wang, Weber, Deininger and Batzer (1995) Alu repeats - a source for the genesis of primate microsatellites. *Genomics*, 29:136-144.

25. Roy-Engel, Salem, Oyeniran, Deininger, Hedges*, et al* (2002) Active alu element "A-tails,": Size does matter. *Genome Research*, 12**:**1333-1344.
26. Calabrese, Durrett and Aquadro (2001) Dynamics of microsatellite divergence under stepwise mutation and proportional slippage/point mutation models. *Genetics*, 159**:**839-852.
27. Johanning, Stevenson, Oyeniran, Gozal, Roy-Engel*, et al* (2003) Potential for retroposition by old Alu subfamilies. *Journal of Molecular Evolution*, 56**:**658-664.
28. Schmid (1998) Does SINE evolution preclude Alu function? *Nucleic Acids Research*, 26**:**4541-4550.
29. Daniels and Deininger (1985) Integration site preferences of the Alu family and similar repetitive DNA-sequences. *Nucleic Acids Research*, 13**:**8939-8954.
30. Arcot, Adamson, Risch, Lafleur, Robichaux*, et al* (1998) High-resolution cartography of recently integrated human chromosome 19-specific Alu fossils. *Journal of Molecular Biology*, 281**:**843-856.
31. Swergold (1990) Identification, characterization, and cell specificity of a human LINE-1 promoter. *Molecular and Cellular Biology*, 10**:**6718-6729.
32. Weiner, Deininger and Efstratiadis (1986) nonviral retroposons - genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. *Annual Review of Biochemistry*, 55**:**631-661.
33. Rogers (1985) The origin and evolution of retroposons. *International Review of Cytology-a Survey of Cell Biology*, 93**:**187-279.
34. Lavie, Maldener, Brouha, Meese and Mayer (2004) The human Ll promoter: Variable transcription initiation sites and a major impact of upstream flanking sequence on promoter activity. *Genome Research*, 14**:**2253-2260.
35. Moran, Holmes, Naas, Deberardinis, Boeke*, et al* (1996) High frequency retrotransposition in cultured mammalian cells. *Cell*, 87**:**917-927.
36. Cost, Feng, Jacquier and Boeke (2002) Human L1 element target-primed reverse transcription in vitro. *Embo Journal*, 21**:**5899-5910.
37. Feng, Moran, Kazazian and Boeke (1996) Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell*, 87**:**905-916.
38. Jurka (1997) Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proceedings of the National Academy of Sciences of the United States of America*, 94**:**1872-1877.
39. Gentles, Kohany and Jurka (2005) Evolutionary diversity and potential recombinogenic role of integration targets of non-LTR retrotransposons. *Molecular Biology and Evolution*, 22**:**1983-1991.
40. Kazazian and Moran (1998) The impact of L1 retrotransposons on the human genome. *Nature Genetics*, 19**:**19-24.
41. Shaikh, Roy, Kim, Batzer and Deininger (1997) cDNAs derived from primary and small cytoplasmic Alu (scAlu) transcripts. *Journal of Molecular Biology*, 271**:**222-234.
42. Aleman, Roy-Engel, Shaikh and Deininger (2000) Cis-acting influences on Alu RNA levels. *Nucleic Acids Research*, 28**:**4755-4761.
43. Shen, Batzer and Deininger (1991) Evolution of the master Alu gene(s). *Journal of Molecular Evolution*, 33**:**311-320.
44. Boissinot, Chevret and Furano (2000) L1 (LINE-1) retrotransposon evolution and amplification in recent human history. *Molecular Biology and Evolution*, 17**:**915-928.
45. Jurka and Smith (1988) A fundamental division in the Alu family of repeated sequences. *Proceedings of the National Academy of Sciences of the United States of America*, 85**:**4775-4778.
46. Kapitonov and Jurka (1996) The age of Alu subfamilies. *Journal of Molecular Evolution*, 42**:**59-65.

47. Salem, Kilroy, Watkins, Jorde and Batzer (2003) Recently integrated Alu elements and human genomic diversity. *Molecular Biology and Evolution*, 20:1349-1361.
48. Batzer, Gudi, Mena, Foltz, Herrera*, et al* (1991) Amplification dynamics of human-specific (HS) Alu family members. *Nucleic Acids Research*, 19:3619-3623.
49. Batzer, Kilroy, Richard, Shaikh, Desselle*, et al* (1990) Structure and variability of recently inserted Alu famili members. *Nucleic Acids Research*, 18:6793-6798.
50. Batzer, Rubin, Hellmannblumberg, Alegriahartman, Leeflang*, et al* (1995) Dispersion and insertion polymorphisms in 2 small subfamilies of recently amplified human Alu repeats. *Journal of Molecular Biology*, 247:418-427.
51. Carroll, Roy-Engel, Nguyen, Salem, Vogel*, et al* (2001) Large-scale analysis of the Alu Ya5 and Yb8 subfamilies and their contribution to human genomic diversity. *Journal of Molecular Biology*, 311:17-40.
52. Jurka (1993) A New subfamily of recently retroposed human Alu repeats. *Nucleic Acids Research*, 21:2252-2252.
53. Leeflang, Chesnokov and Schmid (1993) Mobility of short interspersed repeats within the chimpanzee lineage. *Journal of Molecular Evolution*, 37:566-572.
54. Leeflang, Liu, Chesnokov and Schmid (1993) Phylogenetic isolation of a human Alu flounder gene - drift to new subfamily identity. *Journal of Molecular Evolution*, 37:559-565.
55. Leeflang, Liu, Hashimoto, Choudary and Schmid (1992) Phylogenetic evidence for multiple Alu source genes. *Journal of Molecular Evolution*, 35:7-16.
56. Matera, Hellmann and Schmid (1990) A transpositionally and transcriptionally competent Alu subfamily. *Molecular and Cellular Biology*, 10:5424-5432.
57. Roy, Carroll, Kass, Nguyen, Salem*, et al* (1999) Recently integrated human Alu repeats: finding needles in the haystack. *Genetica*, 107:149-161.
58. Roy-Engel, Carroll, Vogel, Garber, Nguyen*, et al* (2001) Alu insertion polymorphisms for the study of human genomic diversity. *Genetics*, 159:279-290.
59. Deininger and Batzer (1999) Alu repeats and human disease. *Molecular Genetics and Metabolism*, 67:183-193.
60. Batzer, Deininger, Hellmannblumberg, Jurka, Labuda*, et al* (1996) Standardized nomenclature for Alu repeats. *Journal of Molecular Evolution*, 42:3-6.
61. Labuda and Striker (1989) Sequence conservation in Alu evolution. *Nucleic Acids Research*, 17:2477-2491.
62. Zietkiewicz, Richer, Sinnett and Labuda (1998) Monophyletic origin of Alu elements in primates. *Journal of Molecular Evolution*, 47:172-182.
63. Britten, Baron, Stout and Davidson (1988) Sources and evolution of human Alu repeated sequences. *Proceedings of the National Academy of Sciences of the United States of America*, 85:4770-4774.
64. Cordaux, Hedges, Herke and Batzer (2006) Estimating the retrotransposition rate of human Alu elements. *Gene*, 373:134-137.
65. Cordaux and Batzer (2009) The impact of retrotransposons on human genome evolution. *Nature Reviews Genetics*, 10:691-703.
66. Gogvadze and Buzdin (2009) Retroelements and their impact on genome evolution and functioning. *Cellular and Molecular Life Sciences*, 66:3727-3742.
67. Miller and Capy, Mobile genetic elements as natural tools for genome evolution. In *Mobile Genetic Elements: Protocols and Genomic Applications. Volume* 260. Edited by Miller WJCP; 2004: 1-20: *Methods in Molecular Biology*].
68. Callinan, Wang, Herke, Garber, Liang*, et al* (2005) Alu retrotransposition-mediated deletion. *Journal of Molecular Biology*, 348:791-800.

69. Han, Sen, Wang, Callinan, Lee*, et al* (2005) Genomic rearrangements by LINE-1 insertion-mediated deletion in the human and chimpanzee lineages. *Nucleic Acids Research*, 33:4040-4052.
70. Han, Lee, Meyer, Wang, Sen*, et al* (2007) Alu recombination-mediated structural deletions in the chimpanzee genome. *Plos Genetics*, 3:1939-1949.
71. Sen, Han, Wang, Lee, Wang*, et al* (2006) Human genomic deletions mediated by recombination between Alu elements. *American Journal of Human Genetics*, 79:41-53.
72. Bailey, Liu and Eichler (2003) An Alu transposition model for the origin and expansion of human segmental duplications. *American Journal of Human Genetics*, 73:823-834.
73. Jurka, Kohany, Pavlicek, Kapitonov and Jurka (2004) Duplication, coclustering, and selection of human Alu retrotransposons. *Proceedings of the National Academy of Sciences of the United States of America*, 101:1268-1272.
74. Stenger, Lobachev, Gordenin, Darden, Jurka*, et al* (2001) Biased distribution of inverted and direct Alus in the human genome: Implications for insertion, exclusion, and genome stability. *Genome Research*, 11:12-27.
75. Pickeral, Makalowski, Boguski and Boeke (2000) Frequent human genomic DNA transduction driven by LINE-1 retrotransposition. *Genome Research*, 10:411-415.
76. Xing, Wang, Belancio, Cordaux, Deininger*, et al* (2006) Emergence of primate genes by retrotransposon-mediated sequence transduction. *Proceedings of the National Academy of Sciences of the United States of America*, 103:17608-17613.
77. Liu, Zhao, Bailey, Sahinalp, Alkan*, et al* (2003) Analysis of primate genomic variation reveals a repeat-driven expansion of the human genome. *Genome Research*, 13:358-368.
78. Bennettt, Coleman, Tsui, Pittard and Devine (2004) Natural genetic variation caused by transposable elements in humans. *Genetics*, 168:933-951.
79. Xing, Zhang, Han, Salem, Sen*, et al* (2009) Mobile elements create structural variation: Analysis of a complete human genome. *Genome Research*, 19:1516-1526.
80. Mamedov, Shagina, Kurnikova, Novozhilov, Shagin*, et al* (2010) A new set of markers for human identification based on 32 polymorphic Alu insertions. *European Journal of Human Genetics*, 18:808-814.
81. Perna, Batzer, Deininger and Stoneking (1992) Alu insertion polymorphism - a new type of marker for human-population studies. *Human Biology*, 64:641-648.
82. Ryan and Dugaiczyk (1989) Newly arisen DNA repeats in primate phylogeny. *Proceedings of the National Academy of Sciences of the United States of America*, 86:9360-9364.
83. Gasior, Wakeman, Xu and Deininger (2006) The human LINE-1 retrotransposon creates DNA double-strand breaks. *Journal of Molecular Biology*, 357:1383-1393.
84. Jackson and Bartek (2009) The DNA-damage response in human biology and disease. *Nature*, 461:1071-1078.
85. Srikanta, Sen, Huang, Conlin, Rhodes*, et al* (2009) An alternative pathway for Alu retrotransposition suggests a role in DNA double-strand break repair. *Genomics*, 93:205-212.
86. Makalowski, Mitchell and Labuda (1994) Alu sequences in the coding regions of messenger-RNA - source of protein variability. *Trends in Genetics*, 10:188-193.
87. Gerber, Oconnell and Keller (1997) Two forms of human double-stranded RNA-specific editase 1 (hRED1) generated by the insertion of an Alu cassette. *Rna-a Publication of the Rna Society*, 3:453-463.
88. Krogh and Symington (2004) Recombination proteins in yeast. *Annual Review of Genetics*, 38:233-271.
89. Tsubouchi and Roeder (2003) The importance of genetic recombination for fidelity of chromosome pairing in meiosis. *Developmental Cell*, 5:915-925.

90. Walker and Hawley (2000) Hanging on to your homolog: the roles of pairing, synapsis and recombination in the maintenance of homolog adhesion. *Chromosoma*, 109:3-9.
91. Haber (2000) Partners and pathways - repairing a double-strand break. *Trends in Genetics*, 16:259-264.
92. Haber, Ira, Malkova and Sugawara (2004) Repairing a double-strand chromosome break by homologous recombination: revisiting Robin Holliday's model. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences*, 359:79-86.
93. Witherspoon, Watkins, Zhang, Xing, Tolpinrud*, et al* (2009) Alu repeats increase local recombination rates. *Bmc Genomics*, 10.
94. Hsu, Erickson, Zhang, Garver and Heidenreich (2000) Fine linkage and physical mapping suggests cross-over suppression with a retroposon insertion at the npc1 mutation. *Mammalian Genome*, 11:774-778.
95. Rieder, Taylor, Clark and Nickerson (1999) Sequence variation in the human angiotensin converting enzyme. *Nature Genetics*, 22:59-62.
96. Baudat, Manova, Yuen, Jasin and Keeney (2000) Chromosome synapsis defects and sexually dimorphic meiotic progression in mice lacking Spo11. *Molecular Cell*, 6:989-998.
97. Roeder (1997) Meiotic chromosomes: it takes two to tango. *Genes & Development*, 11:2600-2621.
98. Ira, Satory and Haber (2006) Conservative inheritance of newly synthesized DNA in double-strand break-induced gene conversion. *Molecular and Cellular Biology*, 26:9424-9429.
99. Paques and Haber (1999) Multiple pathways of recombination induced by double-strand breaks in Saccharomyces cerevisiae. *Microbiology and Molecular Biology Reviews*, 63:349-+.
100. Slightom, Blechl and Smithies (1980) human-fetal G-gamma-globin AND A-gamma-globin genes - complete nucleotide-sequences suggest that DNA can be exchanged between these duplicated genes. *Cell*, 21:627-638.
101. Chen, Cooper, Chuzhanova, Ferec and Patrinos (2007) Gene conversion: mechanisms, evolution and human disease. *Nature Reviews Genetics*, 8:762-775.
102. Kass, Batzer and Deininger (1995) Gene conversion as a secondary mechanism of short interspersed element (SINE) evolution. *Molecular and Cellular Biology*, 15:19-25.
103. Blanco, Shlumukova, Sargent, Jobling, Affara*, et al* (2000) Divergent outcomes of intrachromosomal recombination on the human Y chromosome: male infertility and recurrent polymorphism. *Journal of Medical Genetics*, 37:752-758.
104. Myers and Mccarroll (2006) New insights into the biological basis of genomic disorders. *Nature Genetics*, 38:1363-1364.
105. Liskay, Letsou and Stachelek (1987) Homology requirement for efficient gene conversion between duplicated chromosomal sequences in mammalian-cells. *Genetics*, 115:161-167.
106. Waldman and Liskay (1988) Dependence of intrachromosomal recombination in mammalian-cells on uninterrupeted homology. *Molecular and Cellular Biology*, 8:5350-5357.
107. Roy, Carroll, Nguyen, Salem, Oldridge*, et al* (2000) Potential gene conversion and source genes for recently integrated Alu elements. *Genome Research*, 10:1485-1495.
108. Zhi (2007) Sequence correlation between neighboring Alu instances suggests post-retrotransposition sequence exchange due to Alu gene conversion. *Gene*, 390:117-121.
109. Kidd, Graves, Newman, Fulton, Hayden*, et al* (2010) A human genome structural variation sequencing resource reveals insights into mutational mechanisms. *Cell*, 143:837-847.
110. Mcvean (2010) What drives recombination hotspots to repeat DNA in humans? *Philosophical Transactions of the Royal Society B-Biological Sciences*, 365:1213-1218.

111. Lee, Han, Meyer, Kim and Batzer (2008) Chromosomal inversions between human and chimpanzee lineages caused by retrotransposons. *Plos One*, 3.
112. Otieno, Carter, Hedges, Walker, Ray*, et al* (2004) Analysis of the human Alu Ya-lineage. *Journal of Molecular Biology*, 342**:**109-118.
113. Styles and Brookfield (2007) Analysis of the features and source gene composition of the AluYg6 subfamily of human retrotransposons. *Bmc Evolutionary Biology*, 7.
114. Roy-Engel, Carroll, El-Sawy, Salem, Garber*, et al* (2002) Non-traditional Alu evolution and primate genomic diversity. *Journal of Molecular Biology*, 316**:**1033-1040.
115. Jurka and Pethiyagoda (1995) Simple repetititve DNA-sequences from primates - compilation and analysis. *Journal of Molecular Evolution*, 40**:**120-126.
116. Kelkar, Tyekucheva, Chiaromonte and Makova (2008) The genome-wide determinants of human and chimpanzee microsatellite evolution. *Genome Research*, 18**:**30-38.
117. Justice, Den, Nguyen, Stoneking, Deininger*, et al* (2001) Phylogenetic analysis of the friedreich ataxia GAA trinucleotide repeat. *Journal of Molecular Evolution*, 52**:**232-238.
118. Kurosaki, Ninokata, Wang and Ueda (2006) Evolutionary scenario for acquisition of CAG repeats in human SCA1 gene. *Gene*, 373**:**23-27.
119. Xing, Wang, Han, Ray, Huang*, et al* (2005) A mobile element based phylogeny of Old World monkeys. *Molecular Phylogenetics and Evolution*, 37**:**872-880.
120. Batzer, Arcot, Phinney, Alegriahartman, Kass*, et al* (1996) Genetic variation of recent Alu insertions in human populations. *Journal of Molecular Evolution*, 42**:**22-29.
121. Salem, Ray, Xing, Callinan, Myers*, et al* (2003) Alu elements and hominid phylogenetics. *Proceedings of the National Academy of Sciences of the United States of America*, 100**:**12787-12791.
122. Ray, Xing, Hedges, Hall, Laborde*, et al* (2005) Alu insertion loci and platyrrhine primate phylogeny. *Molecular Phylogenetics and Evolution*, 35**:**117-126.
123. Ray, Han, Walker and Batzer (2010) Laboratory methods for the analysis of primate mobile elements. *Methods in molecular biology (Clifton, NJ)*, 628**:**153-179.
124. Butler (2006) Genetics and genomics of core short tandem repeat loci used in human identity testing. *Journal of Forensic Sciences*, 51**:**253-265.
125. Giardina, Predazzi, Pietrangeli, Asili, Marsala*, et al* (2007) Frequency assessment of SNPs for forensic identification in different populations. *Forensic science international Genetics*, 1**:**e1-3.
126. Butler, Buel, Crivellente and Mccord (2004) Forensic DNA typing by capillary electrophoresis using the ABI Prism 310 and 3100 genetic analyzers for STR analysis. *Electrophoresis*, 25**:**1397-1412.
127. Ross and Belgrader (1997) Analysis of short tandem repeat polymorphisms in human DNA by matrix-assisted laser desorption ionization mass spectrometry. *Analytical Chemistry*, 69**:**3966-3972.
128. Taylor, Guillen, Nazabal, Fernandez and Silva (2007) Electrophoretic techniques applied to the detection and analysis of the human microsatellite DG10s478. *Journal of biomolecular techniques : JBT*, 18**:**298-305.
129. Witherspoon, Marchani, Watkins, Ostler, Wooding*, et al* (2006) Human population genetic structure and diversity inferred from polymorphic L1 (LINE-1) and Alu insertions. *Human Heredity*, 62**:**30-46.
130. Shewale, Schneida, Wilson, Walker, Batzer*, et al* (2007) Human genomic DNA quantitation system, H-Quant: Development and validation for use in forensic casework. *Journal of Forensic Sciences*, 52**:**364-370.
131. Nicklas and Buel (2003) Development of an Alu-based, real-time PCR method for quantitation of human DNA in forensic samples. *Journal of Forensic Sciences*, 48**:**936-944.

132. Nicklas and Buel (2003) Development of an Alu-based, QSY 7-labeled primer PCR method for quantitation of human DNA in forensic samples. *Journal of Forensic Sciences*, 48:282-291.
133. Nicklas and Buel (2005) An Alu-based, MGB Eclipse (TM) real-time PCR method for quantitation of human DNA in forensic samples. *Journal of Forensic Sciences*, 50:1081-1090.
134. Nicklas and Buel (2006) Simultaneous determination of total human and male DNA using a duplex real-time PCR assay. *Journal of Forensic Sciences*, 51:1005-1015.
135. Sifis, Both and Burgoyne (2002) A more sensitive method for the quantitation of genomic DNA by Alu amplification. *Journal of Forensic Sciences*, 47:589-592.
136. Walker, Kilroy, Xing, Shewale, Sinha*, et al* (2003) Human DNA quantitation using Alu element-based polymerase chain reaction. *Analytical Biochemistry*, 315:122-128.
137. Quental, Azevedo, Rubio, Diogo and Amorim (2009) Molecular mechanisms underlying large genomic deletions in ornithine transcarbamylase (OTC) gene. *Clinical Genetics*, 75:457-464.
138. Tuchman, Mccullough and Yudkoff (2000) The molecular basis of ornithine transcarbamylase deficiency. *European Journal of Pediatrics*, 159:S196-S198.
139. Horwich, Kalousek, Fenton, Pollock and Rosenberg (1986) Targeting of pre-ornithine transcarbamylase to mitochondria - defenition of critical regions and residues in the leader peptide. *Cell*, 44:451-459.
140. Lindgren, Demartinville, Horwich, Rosenberg and Francke (1984) Human ornithine transcabamylase *locus* mapped to band Xp21.1 near the Duchenne muscular-dystrophy *locus*. *Science*, 226:698-700.
141. Horwich, Fenton, Williams, Kalousek, Kraus*, et al* (1984) Structure and expression of a complementary-DNA for the nuclear coded precursor of human mitochondrial ornithine transcarbamylase. *Science*, 224:1068-1074.
142. Plochl, Plochl, Wermuth and Roscher (2001) Variants of inborn errors of metabolism with late onset but nevertheless life threatening course. *Klinische Padiatrie*, 213:261-265.
143. Yamanouchi, Yokoo, Yuhara, Maruyama, Sasaki*, et al* (2002) An autopsy case of ornithine transcarbamylase deficiency. *Brain & Development*, 24:91-94.
144. Rosenberg and Scriver, Disorders of the urea cycle. In *Metabolic Control and Disease.* 1980: 682-687
145. Potter, Hammond, Sim, Green and Wilcken (2001) Ornithine carbamoyltransferase deficiency: Improved sensitivity of testing for protein tolerance in the diagnosis of heterozygotes. *Journal of Inherited Metabolic Disease*, 24:5-14.
146. Sumi, Matsuura, Kidouchi, Togari, Kubota*, et al* (2000) Detection of ornithine transcarbamylase deficiency heterozygotes by measuring of urinary uracil. *International Journal of Molecular Medicine*, 6:177-180.
147. Leibundgut, Wermuth, Colombo and Liechtigallati (1996) Ornithine transcarbamylase deficiency: Characterization of gene mutations and polymorphisms. *Human Mutation*, 8:333-339.
148. Tuchman (1993) Mutations and polymorphisms in the human ornithine transcabamylase gene. *Human Mutation*, 2:174-178.
149. Ellaway, Bennetts, Tuck and Wilcken (1999) Clumsiness, confusion, coma, and valproate. *Lancet*, 353:1408-1408.
150. Honeycutt, Callahan, Rutledge and Evans (1992) Heterozygote ornithine transcarbamylase deficiency presenting as symptomatic hyperammonemia during initiation of valptoate therapy. *Neurology*, 42:666-668.

151. Mccullough, Yudkoff, Batshaw, Wilson, Raper*, et al* (2000) Genotype spectrum of ornithine transcarbamylase deficiency: Correlation with the clinical and biochemical phenotype. *American Journal of Medical Genetics*, 93**:**313-319.
152. Nicolaides, Liebsch, Dale, Leonard and Surtees (2002) Neurological outcome of patients with ornithine carbamoyltransferase deficiency. *Archives of Disease in Childhood*, 86**:**54-56.
153. Burlina, Ogier, Korall and Trefz (2001) Long-term treatment with sodium phenylbutyrate in ornithine transcarbamylase-deficient patients. *Molecular Genetics and Metabolism*, 72**:**351-355.
154. Fox, Hack, Fenton, Golbus, Winter*, et al* (1986) Prenatal-diagnosis of ornithine transcarbamylase deficiency with use of DNA polymorphisms. *New England Journal of Medicine*, 315**:**1205-1208.
155. Fox, Hack, Fenton and Rosenberg (1986) Identification and application of additional restriction-fragment-lenght-polymorphisms at the human ornithine transcarbamylase *locus*. *American Journal of Human Genetics*, 38**:**841-847.
156. Grompe, Caskey and Fenwick (1991) Improved moleular diagnostics for ornithine transcarbamylase deficiency. *American Journal of Human Genetics*, 48**:**212-222.
157. Mcclead, Rozen, Fox, Rosenberg, Menke*, et al* (1986) Clinical-application of DNA analysis in a family with OTC deficiency. *American Journal of Medical Genetics*, 25**:**513-518.
158. Nussbaum, Boggs, Beaudet, Doyle, Potter*, et al* (1986) New mutation and prenatal-diagnosis in ornithine transcarbamylase deficiency. *American Journal of Human Genetics*, 38**:**149-158.
159. Pembrey, Old, Leonard, Rodeck, Warren*, et al* (1985) Prenatal-diagnosis of ornithine carbamoyl transferase deficiency using a gene specific probe. *Journal of Medical Genetics*, 22**:**462-465.
160. Schwartz, Christensen, Christensen, Skovby, Davies*, et al* (1986) Detection and exclusion of carriers of ornithine transcarbamylase deficiency by RFLP analysis. *Clinical Genetics*, 29**:**449-452.
161. Watanabe, Sekizawa, Taguchi, Saito, Yanaihara*, et al* (1998) Prenatal diagnosis of ornithine transcarbamylase deficiency by using a single nucleated erythrocyte from maternal blood. *Human Genetics*, 102**:**611-615.
162. Ray, Gigarel, Bonnefont, Attie, Hamamah*, et al* (2000) First specific preimplantation genetic diagnosis for ornithine transcarbamylase deficiency. *Prenatal Diagnosis*, 20**:**1048-1054.
163. Fox, Rozen, Fenton, Horwich and Rosenberg (1985) Additional restriction fragment lenght polymorphisms (RFLPS) for detection of ornithine transcarbamylase (OTC) deficiency. *Pediatric Research*, 19**:**A247-A247.
164. Hoshide, Matsuura, Komaki, Koike, Ueno*, et al* (1993) Specificity of PCR-SSCP for detection of the mutant ornithine transcarbamylase (*OTC*) gene in patients with OTC deficiency. *Journal of Inherited Metabolic Disease*, 16**:**857-862.
165. Tuchman, Jaleel, Morizono, Sheehy and Lynch (2002) Mutations and polymorphisms in the human ornithine transcarbamylase gene. *Human Mutation*, 19**:**93-107.
166. Engel, Nuoffer, Muehlhausen, Klaus, Largiader*, et al* (2008) Analysis of mRNA transcripts improves the success rate of molecular genetic testing in OTC deficiency. *Molecular Genetics and Metabolism*, 94**:**292-297.
167. Balasubramaniam, Rudduck, Bennetts, Peters, Wilcken*, et al* (2010) Contiguous gene deletion syndrome in a female with ornithine transcarbamylase deficiency. *Molecular Genetics and Metabolism*, 99**:**34-41.
168. Arranz, Madrigal, Riudor, Armengol and Mila (2007) Complete deletion of ornithine transcarbamylase gene confirmed by CGH array of X chromosome. *Journal of Inherited Metabolic Disease*, 30**:**813-813.

169. Shchelochkov, Li, Geraghty, Gallagher, Van Hove*, et al* (2009) High-frequency detection of deletions and variable rearrangements at the ornithine transcarbamylase (OTC) locus by oligonucleotide array CGH. *Molecular Genetics and Metabolism*, 96**:**97-105.
170. Ono, Tsuda, Mouri, Arai, Arinami*, et al* (2010) Contiguous Xp11.4 gene deletion leading to ornithine transcarbamylase deficiency detected by high-density single-nucleotide array. *Clin Pediatr Endocrinol*, 19**:**25-30.
171. Azevedo, Soares, Quental, Vilarinho, Teles*, et al* (2006) Mutational spectrum and linkage disequilibrium patterns at the ornithine transcarbamylase gene (OTC). *Annals of Human Genetics*, 70**:**797-801.
172. Yamaguchi, Brailey, Morizono, Bale and Tuchman (2006) Mutations and polymorphisms in the human ornithine transcarbamylase (OTC) gene. *Human Mutation*, 27**:**626-632.
173. Altschul, Gish and Al. (1990) ENSEMBLE Basic local alignment search tool. *Journal Molecular Biology*, 215**:**403 - 410.
174. Smit, Hubley and Green (2010) RepeatMasker Open-3.0.
175. Kohany, Gentles, Hankus and Jurka (2006) Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *Bmc Bioinformatics*, 7.
176. Drummond, Ashton and Al. (2011) Geneious 5.4 5.4 edition.
177. Rozen and Skaletsky (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods in molecular biology (Clifton, NJ)*, 132**:**365-386.
178. Kibbe (2007) OligoCalc: an online oligonucleotide properties calculator. *Nucleic Acids Research*, 35**:**W43-W46.
179. Altschul, Gish, Miller, Myers and Lipman (1990) Basic local alignment search tool. *Journal of Molecular Biology*, 215**:**403-410.

Appendices

Appendix I: Sequences of the *OTC* Alus

>ALU1
GCTGGGTGCAGTGGCTCATGCTTGTAATCCCTGCACTTTGGGAGGCCGAGGTGGGTGGATCACCTGAGG
TCAGGAGTTCCAGACCAGTTTGGCCAGCATGGCAAAACCCTGTCCCTGTTAAAAATACAAAAAAAGAAA
AAGAATTAGCTGGGCCTGGTGGCATGCACCTGTAGTCCCAGCTACTCAGGAGGCTGAGGCAGAAGAATT
GCTTGAACCTGGGAGGCGGAGGTTGCAGTGAGCCGAGATCGCGCCAATGCACTCCAGCCACCTGGGTGA
CAGAGCGAGACTCTGTCTCAAAAAAAAAAAAAAAAAAAAAAAA

>ALU2
GGCCGGGTGCGGTGGCTCATGCTTGTAATCCCAGCATTTTGGGAGGCTGAGGCAGCTGGATCACCTGAG
GTCAGGAGTTTGAGACCAGCCTGACCAACATGGTGAAACCTCATCTCTACTAAAAATGCAAAAATTAGC
TGGGCATGGTGGCAGACGTCTGTAATCCCAGCTACTCGGGAGGCTGAGGCAGGAGAATCACTTGAACTC
TGGAAGCAGAGGTTGTGGTGAGCTGAGATCGCGCCGTTGCACTCCAGCCTGGGTGACAGAGCAAGACTT
CATCTCAAAAAAAAAAAAAAAAAA

>ALU3
GCCGGGTGCGGTGGCTCATGCCTGTAATCCCAGCACTTTGGGAGGCTGAGGCGGGCAGATCACCTGACG
TCAGGAGTTCGAGACCAGCCTGACCAATATGATGAAACCCCGTCTCTACTAAAAATACAAAAAATTAGC
TGGGCATGGTGGTGAGGGCCTGTAATCCCAGCTATTCGGGAGGCTGAGGCAGGAGAATCACTTGAACCT
GGGAGGCAGAGGTTGCAGTGAGCCAAGATGGTGCCATTGCACTCCAGCCTGGGCAACAAGAGGGAAACT
CCATCTCAGAAAAAAAAAAAAAAGACAAAA

>ALU4
GGCATAGTGGCTCACACCTGTAATCCCAGCACTTTGGGAGGACGAGGCGGGCGGATCACTTGAGGTCAG
GAGTTCTAGACCAGCCTGGCCAACATGGTGAAACCCCGTCTCTACTAAAAATATAAAAATTAGCCGGGC
GTGGTGGCAGGCACCTGTAATCCCAGCTACTCCTTGGGAGGCTGGGGCAGGAGAATCGCTGAAACCCGG
GAGGCAGAGGTTGCAGTAAGCTGAGATCGCACCATTGCACTCTAGCCTGGGTGACAAGAGCGAAACTCT
GCCTCAAAAACAAAAAAAAAAAAAAGAAAGAAA

>ALU5
CTTTGGGAGGCTGAGGCGGGTGGATCACCTGAGGTCAGGAGTTCAAGACCAGCCTGGTCAACATGGTGA
AACTCCATTTCTACTAAAAAACACAAAAATTAGCCAGGCGTGGTGGCAGACGCTTGTAATCCCAGCTAC
TCAGGAGGCTTAGGCAGGAGAATCGCTTGAACCTGGGAGACAGATGTTGCAGTGAGCTGAGATTGCGCC
GCTGCGCTCCAGCCTGGGCGACAGAGCGAGACTCCGTTAAAAAAAAAAAA

>ALU6
GCCGGGCGTGGTGGCTCACGTGCTGTAATTCCAGCAGGTTGGGAGGCTGAGGTGGTTGGATCACCTCAG
GTCAGGAGTTCAAGACCAGCCTGGCCAACATGGTGAAACCCTGTCTCTACCAAAAATACAAAAATTAGC
TGGGCGTGGTGGCACATGCCTGTAGTCCCAGCTACTCGGGAGGCTGAGGCAGGAAAATGTCCTGAAACT
GGGAGGCAGAGGTTGCAGTGAGCCGAGATCACACCACTGCACTCCAGCCTGGGTGACAGAGTGAAACTC
CATCTGAAAAAATAAATAAATAAATAAA

>ALU7
GGCTGGGTGCAGTGGCTCACACCTGTAATCCCAGCACTTTGGGAGGTTGAGGTGGGCAGATCACTTGTG
CTCAGGAGTTCGAGATCAGCATGGCCAACATGATGAAACCCCGTCTCTACTAAAAATACAAAAATTAGC
CAGGTGTGGTGGCAGGTGCCTGTAATCCCAGCTACTCAGGAGGCTGAGGCACAAGAATTGCTTCAACCC
AGGAGGTGGAGGCTGCAATGAGCCGAGATCACGTCACTGTACTCCAGCCTGGGCAACAGAGCAAGACTC
CATCTCAAAAAAAAAAAAA

>ALU8
GAGGCCGAGGCAGGCGGATCACTTGAGCTCAGGAGTTCGAGACCAGCCTGGTCACCATGGAGAAACCCC
GTCTCTAGTAAAAATACAAAAATTAGCCAGGCTTGGTGGCATGTGCTTGTAGTTTCAGCTACTTGGGTG
GCTGAAGCAGGAGAATCGCTTGAACCTGGGAGGCAGAGGCTGCAGTGAACTGAGATTGTGCCACACTCC
AGCCTGGGCCACAGAGTGAGAACCTGTCTCAAAAAAAAAAAAAAGAAAAGAAAAGAA

>ALU9

GGCCGGGCACGGTGGCTCACTCCTGTGATCCCAGCACTTTGGGATGCCAAGGTGGGCGGATCACTTGAG
GTCAAGAGTTTGAGACCAGCCTGGCCAATATGGTAAAACTCCATCTCTACTAAAAATACAAAAATTAGC
CAGGTGTGGTGGTATGCACCTGTAATCCCAGCTACTTGGGAGGCTGAGGCGGGAGAATTCCTTGAACCT
GGGAGGCAGAAGTTGCAGTGAGCCAAGATCACGCCAATGCACTCCAGCCTGGGCAACAGAGCAAGACTC
CATCTCAATAAATAAATAAATAAATAAATAAA

>ALU10
GGCCAGGCGCAGTGGCTCATGCCTGTAATCCCAGCACTTAGGGAGGCCAGGCAGGCAAATCACCTGAGG
TCGGGAGTTCGAGACCAGCCCGACCAACATGAAGAAACCCCGTCTCTACTAAAAATAAAAAATTAGCCG
GGCATGGTGGCACATGCCTGCAATTCCAGCTACTAGGGAGGCTGAGGCAGGAGAATTGCTTGAACCCAG
GAGGCGGAGGTTGCAGTGAGCCAAGATCGTGCCATTGCTGTCCAGCCTGGGCAATAAGAGTGAAACTCC
ATCTGCAAAAAAAGAA

>ALU11
GGCCAGGTGTGGTGGTTCATGCCTTAATCCCAGCACTTTGGGAGGCCAAGGCAGGTGGATCACCTGAGG
TCAGGAGTTCGAGACCAGCCTGGCCAACATGGCAAAACCCCATCTCTACTCAAAATACCAAAAAAAAAA
AAAAAAATTAGCCTGGAGTGGTGGTGGGTGCCTATAATCCCAGCTACTAGAGAGGCTGAGGCAGGAGAA
TTGCTTGAACCTGGGAGATGGAGGTTGCAGTGAGCCAAGATCTTGCCACTGCACTCCAGCCTGGGCAAC
AGAGCAATATTCCATCTCAAAAAAAAAAAA

>ALU12
GGCCGGGCACGGTGGCTCATGCCTGTAATCCCAGCACTTTGGGAGTCCGAGGCAGGTGGATCACGAGGT
CAGGATGTCGAGACCATCCTGGCTAACACAGTGAAACCCCATCTCTACTAAAAATACAAAAAATTAGCC
AGACGTGGTGGCGGGCGCCTGTAGTCCCAGCCAATCGGGAGGCTGAGGCAGGAGAATGGCGTGAACCCG
GGAGGCGGAGCTTGCAGTGAGCCGAGAACGCGCCACTGCTTCCAGCCTGGGCTACAGAGCAAGACTCCA
TCTCAAAAAAAAGAAAAAAAGATAAAAAGAAGAA

>ALU13
GGCCGGGTGCGGTGGCTCACGCCTGTAATCCTAGCACTTTGGGAGGCCAAGGTGGGCGGATCACCTGAG
GTCAGGAGTTTGTGACCAGTCTGGCCAACATGGGGAAACCCATCTCTACTAAAAATACAAAAATTAGCC
TGGCATGGTGGCGGGCGCCTGTAATCCCAGCTACTCGGGAGGCTGAGGCAGGAGGATCACTTGAACTCG
GGAGGTGAAGGTTGCAGTGAGCCAAGCTGCACCACTGCACTGCAGCCAGGGCGAGAGAGTGAGACTTCG
TCTAAAAAAAAGAAA

>ALU14
GAGGCAGGTGGATCACGAGGTCAGGAAATCGAGACCATCCTGGCTAACATGGTGAAATCCCATCTCTAC
TAAAAATACAAAAAATTAGCCAGGCGTGGTGGTGGGCGCCTGTAGTCCCAGCTACTCGGGAGGCTGAGG
CAGGAGGATGGCATGAACCTGGGAGGCAGAGCTTGCAGTGAGATCGCGCCACTGCACTCCACCCTGGGT
GACAGAGCAAGACTCCATCTCCAAAAAGAAAAAAAAA

>ALU15
GGCTGTGTGCAGTGGCTCATACCTATAATCCCAGCACTTTGGGAGGCTGAGGCAGGTGGATCTCCTGAG
GTCAGAAGTTCAAGACCAGCCTGGCCAACATGGCAAAACCCTGTCTCTACTAAAAATACAAAAATTAGC
CGGACGTGGTGGCAGGCGCCTGTAATCCCAGCTAGTCGGGAGGCTGAGGCAGGAGAATCGCTTGAATCT
GGGAGGCAGGGGTTGTAGTGAGCCGAGATCATGCCACTGCACTCCAGCCTGAGCAACAGAGCAAGACTC
TGTCTGAAAAAAATAAAAAATAAAATAAAA

>ALU16
GGCTGGGCGTGGTGGCTCATGCCTATAATCCCAGCACTTTGGGAGGCCGAGGCGGGTGGATCACCTGAG
TTTGGGAGTTTGAGACCAGCCTGACCAACATGGAGAAACCCCGTCTCTACTAAAAATACAAAAATTAGC
CGGGCGTGGTGGCGCATGCCTGTAATCCCAGCTACTCAGGAGGCTGAGGCAGGAGAGTTGCTTAAACTC
GGGAGGCGGAGGTTGCAGTGAGCCGAGATAGCGCCATTGCACTCCAGCCTGGGCAAGAAGAGCAAAACT
CCATCTCAAAAGAAAAGAAAAGACAAAAAAAAAAAAA

>ALU17
GGCTGGGTGCGGTGGCTCACGTCTGTAATCCTAGCACTTTGGGAAGCCGAGGGGCTGGGGGAATGGGGG
TGGGTCACCTGAGGTCAGGACCAGCCTGGCCAACATGGCGAAACCCCATCTCTACTAAAAATACAAAAA
TTAGCTGGGTGTGGTGGCGGGCACCTGTAATCCCAGCTACTCAGGAGGTTGAGGCAGGAGAATCACTTG
AACCCGGGAGGTGGAGGTGGCAGTGAGCTGAGATTGTGCCACTGCACTCCAGCCTGGGCGACAGAGCGG
GACTCTGTCTCAAATAAATAAATAAAATAAAA

>ALU18
GGCCGGGTGTGTTGGCTCACGCCTGTAATCCCAGCACTTTGAGAGGCCGAGGCAGACAGATCACCTGAG
GTCAGGAGTTTGAGACCAGCCTGGCCAACATGGTGAAACCCCATCTCTACTAAAAATAAAATAAAAAAA
ATTCCGCACACCTGTAGTCCCAGCTACTAGGGAGGCTGAGGCAGGAGAATTGCTTGAGCCCAGGAGGCA
GAGGTTGCAGTGAGCCCAGATCCTGCCACTGCACTCCAGCCTGGGCAACAGAGCAAGGCTCTGTTAAAA
AAAAA

>ALU19
GGCCGATGCAGTGGCTCATGCCTGTAATCCCAGCACCTTGGGAGGCCAAGGTGGGCAGATCGCGAGGTC
AGGAGTTTGAGACCAGCCTGACCAACATGGTGAAATCCCGTCTCTACTAAAAATACAAAAATTAGCCAG
GCGTGGTGGCGTGTGCCTGTAATCCCAGCTACTCAGGAGGCAGAGGCAGGAGAATCGCTTGAACCCAGG
AGGTGGAGGTTGTAGTGAGCTGAGGTCACAGCACTGCACCCCAGCCTGGGTGACAGAGTGAGACTCCAT
CTCAAAAAAAAAAAAAAA

>ALU20
TTTGGGAGTCTGAGGCAGGCAGATCACAAGGTCAGGAGTTCGAGACCAGCCTGGCCAACATAGTGAAAC
CCCGTCTCTACTAAAAATACAAAAATTAGCCGGTTGTGGTGGCATGCGACTGTAGTCTCAGCTACTCAG
GAGGCCGAGGCAGGAGAATCACTTGAACCAGGGAGGTGGAGATTGCAGTGAGCCAAGACCATGCCACTA
CACTCTAGCCTGGGTGACAGAGTGAGATTTCGTCTCAAAAAACAAAAAAAGAA

>ALU21
GGCAGGGTGCAGTGGCTCATGCCTGTAATCCCAGCACTTTGGGAAGCTGAGGCAGGCGGATCACCTGAG
GTCAGGAGTTCGAGACCAGCTTGGCTAACATGGTGAAACCCCATTTCTACTAAAAATACAAAAAAAAAA
AAAATTAGCCGGGCGTGGTGGCATGTGCCTATAATCCCAGCTACTCAGGAGGCTGAGGCAGGAGAATCA
CTTGAATCCGGGAGGCAGAGGCTGCATTGCACTCCAGCTTGGGCAACAAGAGCAAAACTCCATCTCAAA
AAAAAAAAAA

>ALU22
GCCAAGTGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGAGAGTCTGAGACGGGTGATCACCTGAGGT
CAGGAGTTTGAAACCAGCCAGGCCAACATGGTGAAACCTCGTCTTTACTAATAACACAAAAATTAGCCG
GGCGTAGTGGCGCATGCCTCTAATCCAAGCTACTTGGTAGACTGAGGCCGGATAATTGCTTGAACCTGG
GAGGGAGAGGTTGCAGTGAGCCGAGATCATGCCACTGCACTCCAGCCTGGGCGACAGAGCGAGACTCCG
TCTGAAAAAAAAAAAA

>ALU23
GCCAGGTGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGGGGCCGAGGTGGGCGGATCACCTGAGG
TCAGGAGTTTGAGACCAGCCTGGCCAACATGACAAAACCCTGTCTCTACTAAAAATACAAAAATTAGCT
GGGTGTGGTGGCGAGCCTGTAATCCTAGCTACTTGGGAGGCTAAGGTGAGAGAATCGATTGTACCTGGG
AGGAGGGGGTTGCAGTGAGTTGAGATCACGCCACTGCCCTCCAGCCAGGGAGACAGAGCAAGACTCCAA
CTCAAAAAAAACAAACAAA

>ALU24
GGCCAGGTGCAGTGGCTCATTCCTGTAATCCCAGCACTTTGGGAGACCAAGGTGGGCGGATCACCTGAG
GTCAGAAGTTCGAGACCAGCCTGGTCAACATGATGAAACCCCATGTCTACTAAAAACACAAAAATTAGC
CAGGCGTGGCACACACTTGTAATCCCAGCTACTCCGGAGGCTGAGGCACAAGAATCGCTTGAACCCGGG
AGGCGGAGGTTGCAGTGAGTTGAGATCATGCCACTGCACTCCAGTCTGGGTGACAGAGTGAGACTCTGT
CTCAAATAATAATAATAACAAAAA

>ALU25
GCCAGACACGGTGGCTCATGCCTGTAATCCCAGCACTTTGAGAGGCGAGGCGGGCGGATCATTTGAGGT
CAGGAGTTTAAAACCAGCCTGGCCAATATGGTGAAATCCCGTCTCTACTAAAAATACAAAAATTCACTG
GGCATGGTGGCACATGCCTGTAGTCCCAGCTACTCAGGAAGCTGAGGCAGGAGAATTGCTTGAGCTCTG
GAGGCGGAGGTTGCAGTCAGCCGAGATTTTGCCACTGCACTCCAGCCTGGGTGACAGAGCGAGACTCTG
TCTCAAAAAAAAAAAAAAA

>ALU26
GGCCTGGCACGGTGGCTCACGCCTGTAATCCCAGTGCTTTTCGAGGCTGAGGTGGGTGGATCATTTGAG
GTCAGGAGTTCGAGACCAGCCTGGGCAACATGATGAAACCCCCCTACCCGCCACTAAAAATACAAAAAT
TAGCTATGCATGGTGTCACCTGCCTGTAATCCCAGCTACTTGTGAGGCTGAGGTAGCAAAATTGCTTGA
ACCCGGGAGGCAGAGGTTGCAGTGAGCCGAGATCACACCACTGCACTCCAGCCTGGGCAACAGAGCGAG
ACTCTATCTCAAAAAAAAACAAAAAAAAAAAA

>ALU27
GGCCGGCTGTGGTGGCTCACACCTATAATCCCAGCACTTTGGGAGGCCGAGGCAGGTGGATCTCCTGAG
GTCAGGAGTTCGAGACCAGCCTGGCCAATATGGTGAAGCCCTGCCTCTACTAAAAATACAAAAATTAGC
CGGGCACAGTGGCGGGCACCTGTAATCCCAGCTACTTGGGAGGCTGAGGCAGGAGAATCACTTGAACCC
AGGAGGCGGAGGTTGCAGTGAGCCGAGATCGTGCCATTGTACTCCAACCTGGGTGACAGAGCGAGACTC
TGTCTCAAAAAATAAAAATAATAATAAAA

>ALU28
GGCCAGGCGCACTGGCTCATGCCTATAATCCCAGCACTTTGGGAGGCTGAGACAGGGGGATTACTTGAG
GTCAGGAGTTTGAGACCAGCCTGGCCAACATGGTAAAACTCTGTCTCTACTAAAAATACAAAAATATTA
GCCTGGCATGGTGGCGCACATCTGTAATCCCAGCTACTCAGGAGGCTGAGGCAGGAGAATTGCTTGAAC
CTGGGAGGCAGAGGTTGCAGTGAGCCAAGATCGTGCCACTGCACTCCAGCCTGGGCGACAGAGCAAGAC
TCCGTCTCAAAACAAACAAACAAGCAAACAAACAAAAAAA