J Mol Evol (1995) 41:974–978

JOURNAL OF MOLECULAR EVOLUTION

Repetitive DNA Sequences Located in the Central Region of the Human *mdr1* (Multidrug Resistance) Gene May Account for a Gene Fusion Event During Its Evolution

Marc Pauly,[1] Isabelle Kayser,[1] Martine Schmitz,[1] Fernand Ries,[2] François Hentges,[2] Mario Dicato[2]

[1] Laboratoire de Recherche sur le Cancer et les Maladies du Sang, Bâtiment des Sciences, Centre Universitaire de Luxembourg, Avenue de la Faïencerie 162A, L-1511 Luxembourg, Grand-Duchy of Luxembourg
[2] Service de Médecine Interne, Centre Hospitalier de Luxembourg, Rue Barblé 4, L-1210 Luxembourg, Grand-Duchy of Luxembourg

Received: 3 March 1995 / Accepted: 29 May 1995

Abstract. The *mdr1* gene, first member of the human multidrug-resistance gene family, is a major gene involved in cellular resistance to several drugs used in anticancer chemotherapy. Its product, the drug-excreting P-glycoprotein, shows a bipartite structure formed by two similar adjacent halves. According to one hypothesis, the fusion of two related ancestral genes during evolution could have resulted in this structure. The DNA sequence analysis of the introns located in the region connecting the two halves of the human *mdr1* gene revealed a highly conserved poly(CA) · poly (TG) sequence in intron 15 and repeated sequences of the *Alu* family in introns 14 and 17. These repeated sequences most likely represent "molecular fossils" of ancient DNA elements which were involved in such a recombination event.

Key words: Gene evolution — Gene fusion — Multidrug resistance — Human *mdr1* gene — P-glycoprotein — Intron — DNA sequence — *Alu* repeat — poly(CA) · poly(TG) microsatellite — Conservation

Introduction

The multiple-drug-resistance (MDR) phenotype, characterized as a cellular resistance to several, even unrelated cytotoxic substances simultaneously, is frequently associated with the failure of anticancer chemotherapy. The *mdr1* gene is a major gene involved in this phenomenon and encodes the 170-kd membrane P-glycoprotein (*Pgp*) (Ueda et al. 1987). The *Pgp* sequence of 1,280 amino acids is organized in a bipartite structure of two similar adjacent halves, each encompassing a hydrophobic region with six predicted transmembrane domains and a hydrophilic region with a nucleotide-binding domain (Chen et al. 1986). Because of the similarity shared with bacterial transport proteins, the *Pgp* is believed to be involved as a drug-efflux pump in an active transmembrane transport driven by the hydrolysis of adenosine triphosphate (ATP) at the nucleotide-binding sites (Chen et al. 1986). The natural physiological role of the *Pgp* could therefore consist of a cellular detoxification mechanism excreting different toxins with a broad substrate specificity. Under the selective pressure of cytotoxic drugs, the MDR phenotype often appears in surviving cells after *Pgp* overproduction, resulting from either *mdr1* gene amplification or activation (reviewed in Gottesman and Pastan 1993). The *mdr1* gene represents the first of two homologous members in the human *mdr* gene family localized on chromosome 7q21.1 (Callen et al. 1987). Both members could have originated from one common precursor gene by duplication (Chen et al. 1990). Similar *mdr* gene families with three members were found in mouse, hamster, and rat (reviewed in Gottesman and Pastan 1993). It is tempting to believe

Correspondence to: M. Pauly

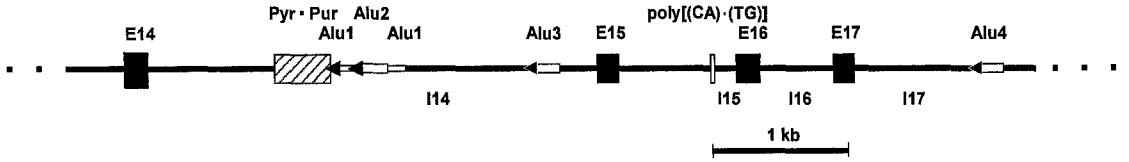

Fig. 1. Structural organization in the connecting region of the human *mdr1* gene. Exons and introns are represented by *full black boxes and lines*, respectively. Exon sizes were taken from Chen et al. (1990). *Arrows* indicate the location and orientation of the *Alu* repeated sequences. The location of the poly(CA) · poly(TG) sequence is represented by a *vertical white bar* and that of the homopyrimidine · homopurine sequence (Pyr · Pur) by a *hatched box*. The scheme is drawn to scale as indicated below.

that all *mdr* families derived from the precursor gene at the beginning of mammal appearance.

Since most intron locations in the two halves of the human *mdr1* gene are different, a hypothetical evolutionary mechanism has been proposed in which the bipartite structure of the precursor *Pgp* gene arose from the fusion of two related but independent genes rather than from the duplication of a single gene (Chen et al. 1990). In order to identify specific DNA sequences as potential "molecular fossils" accounting for an ancient gene fusion event, we analyzed the introns located in the region connecting the two similar halves of the human *mdr1* gene.

Materials and Methods

Gene Source and Cloning. A human (*Homo sapiens sapiens,* Caucasian race, male adult) leucocyte genomic DNA library was screened by hybridization with human *mdr1* cDNA fragment 5A (Ueda et al. 1987) as a probe to isolate a positive phage λ Charon40 clone with a human *Sau*3AI-*Bam*HI insert of about 17 kb. This insert was digested by restriction endonuclease *Xba*I (Life Technologies, Gaithersburg, MD, USA) and the resulting fragments were subcloned into plasmid vector pUC19 (Pharmacia, Milwaukee, WI, USA) for sequencing. DNA extraction and gene cloning were as described for standard methods (Sambrook et al. 1989).

DNA Sequencing and Sequence Analysis. DNA sequencing was performed using the dideoxynucleotide method (Sanger et al. 1977) on double-stranded DNA templates with T7 DNA polymerase (Pharmacia) and synthetic primers. DNA sequence data were analyzed by the Hibio DNAsis software (Hitachi, Yokohama, Japan).

Length Polymorphism Analysis. In the case of the poly(CA) · poly(TG) sequence in intron 15 of human *mdr1* gene, a 310-bp region covering this sequence was amplified by the polymerase chain reaction (PCR) performed in 25 μl on 100 ng of genomic DNA (or 50 pg of plasmid DNA) using 0.75 units of native Taq DNA polymerase (Perkin Elmer, Roche Molecular Systems, Branchburg, NJ, USA) with 12.5 pmol sense (CA-strand) primer 5′-AGTCAAAACACCCACAAAAAAT-3′ and 12.5 pmol antisense (TG-strand) primer 5′-CAACTTCATTTCCTGCTGTCTA-3′, 5 μCi α-^{32}P dATP (800 Ci/mmol) in 200 μM each of dATP, dCTP, dGTP, and dTTP, 10 mM Tris-HCl pH 8.3, 50 mM KCl, and 1.5 mM $MgCl_2$. The initial denaturation was performed for 5 min at 95°C; the amplification encompassed 30 cycles of 1 min at 95°C, 2 min at 59°C, and 3 min at 72°C and was followed by a final extension step of 10 min at 72°C. PCR products were then cleaved by staggered-cutting *Pst*I endonuclease (Life Technologies) and electrophoresed on a 4% denaturing polyacrylamide gel to separate the sense (CA) and antisense (TG) strands. The strand fragments were then visualized by autoradiography and the fragment size determined by comparison to a phage M13 sequencing ladder run in parallel as a size standard. Direct sequencing of the PCR products was performed by the Sanger method as described above, after electroeluting the samples out of the gel.

Results

We analyzed the DNA nucleotide sequences of introns 14, 15, 16, and 17 in the region connecting the two similar halves of the human *mdr1* gene. In intron 14 (total size: 3,316 bp, EMBL Data Library Accession Number: X78081), we previously identified a homopyrimidine · homopurine sequence (Pyr · Pur) (total size: 410 bp, nucleotide position (starting at the 5′-end of the intron with +1): 935–1344), followed by a cluster of inverted *Alu* repeated sequences: *Alu 1* (total size: 269 bp, position: 1345–1479 and 1767–1900), interrupted by the insertion of *Alu 2* (total size: 287 bp, position: 1480–1766), and *Alu 3* (total size: 278 bp, position: 2757–3034) (Pauly et al. 1995; Fig. 1).

Intron 15 (total size: 863 bp, EMBL Data Library Accession Number: X69317), representing the "hinge" between the two halves of the gene, harbors an almost homogenous tandemly repeated sequence of the dinucleotide d(CA · TG) (total size: 32 bp, position 687–718) (Figs. 1, 2). While most of such DNA sequences, also known as microsatellites, exhibit a high degree of length polymorphism consisting in a variable number of tandem repeats (VNTR) from one individual to another in one species (Weber and May 1989), we showed that the poly(CA) · poly(TG) in intron 15 of the human *mdr1* gene is very stable in length and has thus been well conserved during evolution: length polymorphism analysis performed on 110 unrelated Caucasian individuals only revealed two cases of a heterozygous deletion of 5 dinucleotides at the 3′-end of the repeated sequence (position 709–718), as confirmed by direct sequencing (Genome Data Base identification number: 42187, Fig. 3). The high length stability in the case of this imperfect dinucleotide repeated sequence is consistent with the existence of a correlation between the level of length polymorphism and the structural heterogeneity of a short tandemly repeated sequence: the interspecific comparison of the polymorphism and the sequence of evolutionarily conserved microsatellites between several mammalian species revealed a decrease in the number of allele sizes

```
571 TGAGTCAAAACACCCACAAAAAATGGCGTGAGAAGATAAGATACATGGAA
    ----------------------

621 TGAATGCCAGATTATTTCAGTATATGCAGTGGGAATTCTTCTTTGGTGGT

671 TTTCATTTTCAAATATCACACACACACACAGACACACACAGACACACATA
                ==============================

                                  ↓
721 AAACAACACAGCAGATTAGCTTTATCCTTTTCTGCAGTTACTAAACAAAT
                                  ******

771 TGCTGTTTTCCTTGATAGCTGACATTCAGTGATTAGCTTTCATTGGTTAA

                                                  1888
                                                   |
821 CACACAGCCTAATGAGCTTTTGCATATTTCTTATTTATTTTAGACAGCAG
                                                  ----------

    GAAATGAAGTTG
    ------------
```

Fig. 2. Structural organization and DNA nucleotide sequence of the coding strand in the poly(CA) · poly(TG) sequence region in intron 15 of the human *mdr1* gene. The poly(CA) · poly(TG) sequence is underlined with a *double broken line,* the hybridization sites of the PCR primers used in length polymorphism analysis with *single broken lines,* the *Pst*I restriction site with *stars,* and the cleavage position of the *Pst*I endonuclease is indicated by a *vertical arrow.* The nucleotide position is shown at the *left-hand margin* and starts with +1 at the 5′-end of intron 15. Number 1888 refers to the cDNA residue position and corresponds to the start of exon 16 (Chen et al. 1990).

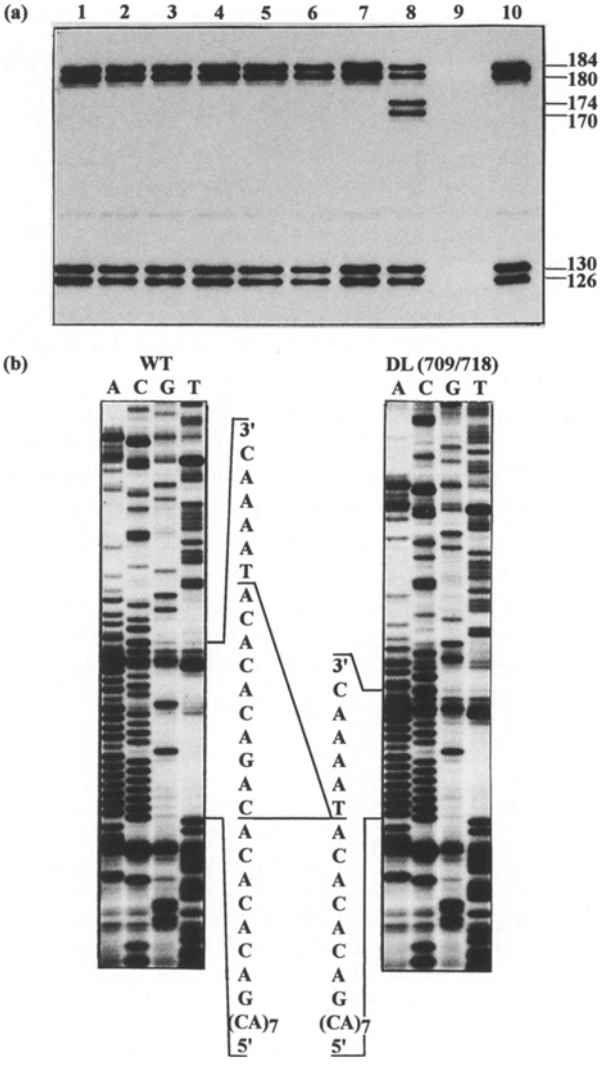

Fig. 3. Length polymorphism analysis of the poly(CA) · poly(TG) sequence in intron 15 of the human *mdr1* gene. **a** Size determination of PCR products by denaturing gel electrophoresis. PCR products were cleaved and analyzed according to Materials and Methods. *Lanes 1–7:* Unrelated Caucasian individuals homozygous for the most frequent allele Z (genotype (Z, Z), tandem repeat number: 16, PCR product size: 310 bp). *Lane 8:* Caucasian individual heterozygous for the 709/718 deletion (genotype (Z, Z-10), tandem repeat numbers: 16, 11 (deleted), PCR product sizes: 310, 300 bp (dl.)). *Lane 9:* Negative control with water. *Lane 10:* Positive control with a recombinant pUC19 plasmid harboring intron 15 of the human *mdr1* gene. The fragment size (bp) is indicated at the *right-hand margin.* The *Pst*I-fragments covering the poly(CA) · poly(TG) region have a size of 184, 174 bp (dl.) for the sense (CA) strand and of 180, 170 bp (dl.) for the antisense (TG) strand, respectively. **b** Nucleotide sequence determination of PCR products by direct dideoxynucleotide sequencing. PCR products were purified and sequenced according to Materials and Methods. *WT:* PCR product corresponding to allele Z with a "wild type" sequence. *DL:* PCR product corresponding to allele Z-10 with the 709/718 deleted sequence. The nucleotide sequence from position 687 to 724 (cf. Fig. 2), as determined in each case, is represented in a *vertical orientation in the center.* The accurate position of the 709/718 deletion is indicated by *full lines.*

at these chromosomal loci linked to interruptions in their repeated sequence (Pépin et al. 1995). Point mutations, as well as short insertions and deletions in the microsatellite sequence, could possibly prevent a replication slippage recognized as the predominant molecular mechanism generating new allele sizes at microsatellite loci not only *in vitro,* but also *in vivo* (Pépin et al. 1995).

In intron 16 (total size: 547 bp, EMBL Data Library accession number: X69318), there was no particular DNA sequence, except one palindromic sequence 5′-$(T)_8(A)_8$-3′ (total size: 20 bp, position 325–344).

However, in the 5′ part of intron 17 (EMBL Data Library accession number: X83289), we discovered another *Alu* repeated sequence, which we named *Alu 4* (total size: 255 bp, position 841–1095, Figs. 1, 4). *Alu 4* shows a high degree of similarity (78%) to the *Alu* consensus sequence described elsewhere (Britten et al. 1988). The structure exhibits the organization of a typical *Alu* element, the latter encompassing a head-to-tail tandem dimer of about 300 bp flanked by two direct repeated sequences (Zuckerkandl et al. 1989, Fig. 4). However, the usual 3′-terminal polyadenine tail frequently observed in the case of classic *Alu* repeated sequences and which could be expected, in this case of inverted orientation, as a polythymine sequence at the 5′-end, has been deleted and/or merged into the T-rich direct repeats.

Discussion

The gene fusion hypothesis suggests that the *mdr* precursor gene originated from the fusion of two related, but independent ancestral genes during evolution (Chen et al. 1990; and Introduction). This recombination event could have occurred in a common ancestor of the modern mammals. The repetitive DNA sequences, namely the poly(CA) · poly(TG) sequence in intron 15 and the *Alu* repeated sequences in introns 14 and 17 of the region

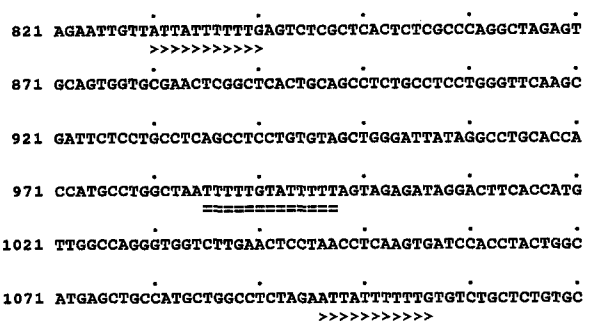

Fig. 4. Structural organization and DNA nucleotide sequence of the coding strand of *Alu4* in intron 17 of the human *mdr1* gene. The *Alu* repeated sequence named *Alu4* is flanked by direct repeated sequences indicated by *arrowheads*. The head-to-tail tandem-arranged monomers of the *Alu* element are separated by a T-rich spacer sequence which is underlined by a *broken line*. The nucleotide position is shown at the *left-hand margin* and starts with +1 at the 5′-end of intron 17.

connecting the two similar halves of the human *mdr1* gene, most likely represent "molecular fossils" of ancient DNA elements which could have been actively involved in the underlying mechanism.

A DNA looping or wrapping event pairing the chromosomal regions containing the related ancestral genes most probably represented the first main step in this recombination process. Repetitive DNA sequences of the alternating pyrimidine-purine $(CA \cdot TG)_n$ motif with a number n of tandem repeats up to 30 have been demonstrated to spontaneously and stably associate with each other into four-stranded DNA complexes (Gaillard and Strauss 1994). They are dispersed throughout the whole mammalian genome with an estimated copy number of about 10^5 and frequency of once every 30–60 kb (Weber and May 1989). In addition, DNA strands of different *Alu* repeated elements can temporarily hybridize to another on the basis of their sequence similarity, and in this way, even distant genomic regions can be brought into close vicinity. *Alu* repeated sequences are even more frequent than poly(CA) · poly(TG) sequences, with an estimated copy number of at least 4.5×10^5 per genome (Zuckerkandl et al. 1989), giving a minimal frequency of once every 6–7 kb. The number of these repeated sequences is thus high enough to allow an association at any genomic location between large regions containing at least one entire average-sized gene. Both the poly(CA) · poly(TG) sequence located in intron 15 and the *Alu* repeated sequences in introns 14 and 17 could thus have induced or at least facilitated DNA pairing.

Genetic exchange between the paired DNA strands logically constituted the second main step in recombination. Poly(CA) · poly(TG) sequences were shown to adopt a left-handed Z-DNA conformation in the appropriate conditions *in vivo* (Lancillotti et al. 1987), to interact with eucaryotic recombination enzymes (Fishel et al. 1988; Gaillard and Strauss 1994), and to enhance reciprocal exchange between similarly structured DNA strands at crossing-over points within their vicinity during gene conversion events (Bullock et al. 1986; Treco and Arnheim 1986; Wahls et al. 1990). Maximum stimulation of recombination was observed when a poly(CA) · poly(TG) sequence was present in both of the paired DNA molecules, but the presence of a single sequence of this type, like in the case of intron 15 of the human *mdr1* gene, is enough to account for a recombination event. This was shown in experiments with DNA substrates where only one contained a $(CA)_{30}$ sequence and preferentially acted as the recipient of a DNA fragment (Wahls et al. 1990). The ancestor of the evolutionarily highly conserved poly(CA) · poly(TG) sequence located in intron 15, representing the "hinge" between the two parts of the human *mdr1* gene, appears as the only potential DNA sequence exhibiting the appropriate properties to have likewise promoted a gene fusion event connecting two similar, but independent genes into the precursor *Pgp* gene.

Furthermore, DNA sequences of the homopyrimidine · homopurine (Pyr · Pur or Y · R) type have been shown to be able to adopt a nonconventional spatial structure different from the classic B-helix as defined by Watson and Crick and are therefore supposed to disable the normal chromatin organization in the region where they are located (Liu and Chan 1990). Such genomic regions are likely to be free of histones and readily accessible to enzymes involved in retroposition and recombination processes, favoring not only the generation of retroposon clusters like those found in the human *nucleophosmin* (B23) gene (Liu and Chan 1990) and in the human *mdr1* gene (Pauly et al. 1995), but also the appearance of preferential recombination sites. The presence of a homopyrimidine · homopurine sequence, such as the one we identified in intron 14 of the human *mdr1* gene (Pauly et al. 1995), therefore is yet another argument in favor of the gene fusion hypothesis.

Acknowledgments. We are thankful to Prof. J.-M. Garnier, Prof. P. Chambon (Inst. Chim. Biol.), Dr. A. Steinmetz (IBMP), and Prof. E. Westhof (IBMC, Strasbourg, France) for the gift of the human genomic library and helpful advice. Further, we would like to thank Dr. I. Pastan (NIH, Bethesda, MD, USA) and Prof. P. Reizenstein† (Karolinska Inst., Stockholm, Sweden) for probe 5A, Dr. S. Chevillard, Dr. M. Dutreix (Inst. Curie), and Dr. F. Strauss (Inst. J. Monod, Paris, France) for useful discussions, and Prof. S. Demczuk (Karolinska Inst.) for critical reading of the manuscript. This research project was supported by the Centre de Recherche Public (CRP) de la Santé and by the Fondation de Recherche Cancer et Sang of Luxembourg.

References

Britten RJ, Baron WF, Stout DB, Davidson EH (1988) Sources and evolution of human *Alu* repeated sequences. Proc Natl Acad Sci USA 85:4770–4774

Bullock P, Miller J, Botchan M (1986) Effects of poly-[d(pGpT) · d(pApC)] and poly[d(pCpG) · d(pCpG)] repeats on homologous recombination in somatic cells. Mol Cell Biol 6:3948–3953

Callen DF, Baker E, Simmers RN, Seshardri R, Roninson IB (1987) Localization of the human multiple drug resistance gene, *MDR1,* to 7q21.1. Hum Genet 77:142–144

Chen CJ, Chin JE, Ueda K, Clark DP, Pastan I, Gottesman MM, Roninson IB (1986) Internal duplication and homology with bacterial transport proteins in the *mdr1* (P-glycoprotein) gene from multidrug-resistant human cells. Cell 47:381–389

Chen CJ, Clark D, Ueda K, Pastan I, Gottesman MM, Roninson IB (1990) Genomic organization of the human multidrug resistance (*MDR1*) gene and origin of P-glycoproteins. J Biol Chem 265:506–514

Fishel RA, Detmer K, Rich A (1988) Identification of homologous pairing and strand-exchange activity from a human tumor cell line based on Z-DNA affinity chromatography. Proc Natl Acad Sci USA 85:36–40

Gaillard C, Strauss F (1994) Association of poly(CA) · poly(TG) DNA fragments into four-stranded complexes bound by *HMG1* and *2.* Science 264:433–436

Gottesman MM, Pastan I (1993) Biochemistry of multidrug resistance mediated by the multidrug transporter. Annu Rev Biochem 62:385–427

Lancillotti F, Lopez MC, Arias P, Alonso C (1987) Z-DNA in transcriptionally active chromosomes. Proc Natl Acad Sci USA 84: 1560–1564

Liu QR, Chan PK (1990) Identification of a long stretch of homopurine · homopyrimidine sequence in a cluster of retroposons in the human genome. J Mol Biol 212:453–459

Pauly M, Kayser I, Schmitz M, Ries F, Hentges F, Dicato M (1995) The human *mdr1* (multidrug-resistance) gene harbours a long homopyrimidine · homopurine sequence next to a cluster of Alu repeated sequences in intron 14. Gene 153:299–300

Pépin L, Amigues Y, Lépingle A, Berthier JL, Bensaid A, Vaiman D (1995) Sequence conservation of microsatellites between *Bos taurus* (cattle), *Capra hircus* (goat) and related species. Examples of use in parentage testing and phylogeny analysis. Heredity 74:53–61

Sambrook J, Fritsch EF, Maniatis T (1989) Molecular cloning. A laboratory manual, 2nd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74:5463–5467

Treco D, Arnheim N (1986) The evolutionarily conserved repetitive sequence d (TG · AC)$_n$ promotes reciprocal exchange and generates unusual recombinant tetrads during yeast meiosis. Mol Cell Biol 6:3934–3947

Ueda K, Cardarelli C, Gottesman MM, Pastan I (1987) Expression of a full-length cDNA for the human "*MDR1*" gene confers resistance to colchicine, doxorubicin, and vinblastine. Proc Natl Acad Sci USA 84:3004–3008

Wahls WP, Wallace LJ, Moore PD (1990) The Z-DNA motif d(TG)$_{30}$ promotes reception of information during gene conversion events while stimulating homologous recombination in human cells in culture. Mol Cell Biol 10:785–793

Weber JL, May PE (1989) Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. Am J Hum Genet 44:388–396

Zuckerkandl E, Latter G, Jurka J (1989) Maintenance of function without selection: *Alu* sequences as "cheap genes." J Mol Evol 29: 504–512