THYROID
Volume 13, Number 4, 2003

Identification and Characterization of a Novel Large Insertion/Deletion Polymorphism of 1464 Base Pair in the Human Thyroglobulin Gene

Christian M. Moya, Viviana Varela, Carina M. Rivolta, Fernando M. Mendive, and Héctor M. Targovnik

We identified a novel large insertion/deletion (Indel) polymorphism of 1464 bp localized in intron 18 of the human thyroglobulin gene. Data from sequence showed a high A+T content (62%), two 17-bp long motif repeats, and three different types of 10-bp long palindromic sequences. The comparison between these 1464 bp and sequences deposited in National Center for Biotechnology Information (NCBI)/GenBank database exhibit a nonsignificant degree of homology with any previously described sequences. The long polymerase chain reaction (PCR) method was used to amplify the genomic DNA region containing intron 17/exon 18/intron 18/exon 19/intron 19 by primers situated in the introns 17 and 19. The amplification generates two fragments of 3.5 and 5.0 kb that correspond to the exclusion or inclusion of a 1464-bp segment, respectively. Both variants are thus widely represented in the human population; giving allele frequencies of 0.56 (insertion) and 0.44 (deletion). Finally, the polymorphism was confirmed by sequence analysis of the 5.0- and 3.5-kb amplified fragments.

Introduction

THYROGLOBULIN (Tg), a large dimeric glycoprotein, is the thyroid hormone's precursor and storage form of iodine and inactive hormone (1). Recently, the complete structure of human Tg gene has been determined (2–9). It is coded by a single-copy gene, 270 kb long, that maps on chromosome 8q24 and contains 8.5 kb of coding sequence divided into 48 exons. The exons' sizes range from 63 to 1101 nucleotides, each of which is separated by introns varying in size up to 64 kb. The preprotein monomer is composed of a 19 amino acid signal peptide followed by a 2749-residue polypeptide. Eighty percent of the monomeric primary structure is characterized by the presence of three types of repetitives units. Five hormonogenic acceptor tyrosines have been identified and localized at positions 5, 1291, 2554, 2568, and 2747. The efficient coupling of the hormone precursors mono and diiodotyrosine to form triiodothyronine (T_3) and thyroxine (T_4) depends on the integrity of the Tg structure. Several mutations of the Tg gene have been reported and are associated with congenital goiter and variable degrees of hypothyroidism (10–16) or endemic and nonendemic simple goiter (17–19). Also, various single nucleotide polymorphisms (SNP) have been defined recently in this gene (1). The availability of highly informative polymorphic markers will allow indirect disease diagnosis by genetic linkage studies in those cases without an identifiable mutation or for rapid identification of affected newborns or gene carriers in families with Tg mutation.

In the present paper, we report the identification of a novel large insertion/deletion (Indel) polymorphism of 1464 bp localized in intron 18 of the human Tg gene. We have also established allele frequencies and structural characteristics.

Materials and Methods

Preparation of λ phage DNA

Bacteriophage DNA was prepared with the Wizard λ preparations DNA purification system (Promega, Madison, WI). After elution from the minicolumn, DNA was extracted twice with phenol-chloroform, salt concentration was adjusted to 2 mol/L with amonium acetate and the DNA was precipitated with ethanol.

DNA sequencing of recombinant phage

The sequences were determined by *Taq* polymerase-based chain terminator method (fmol, Promega) from λ phage clone DNA. Oligonucleotide sequences are shown in Figure 1. The results were analyzed using the PC gene (Intellige-

Laboratorio de Biología Molecular, Cátedra de Genética y Biología Molecular, Facultad de Farmacia y Bioquímica. Universidad de Buenos Aires, Buenos Aires, Argentina.

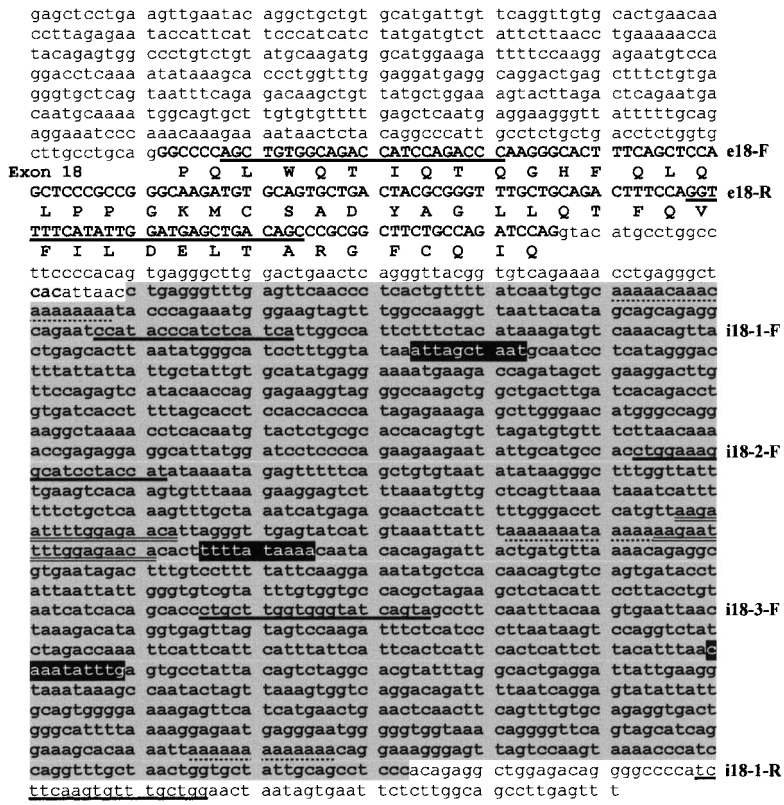

FIG. 1. Sequence data of the exon 18, 3′ end of intron 17 and the insertion/deletion polymorphic region of intron 18 of the thyroglobulin gen, from λ dash 62. The exonic sequences are indicated by capital letters and the intronic sequences by lower-case letters. The amino acid sequences are represented by the one-letter code. The exon maps between positions 3848 and 4002 of the thyroglobulin (Tg) mRNA. Shaded area indicates the 1464-bp polymorphic insertion/deletion region. The positions of the sequence primers are underlined. The two 17-bp long repeat motifs are double underlined, the three palindromic sequences are indicated by darker areas and the three short $(A)_n$ repeat traits by dotted lines.

netics, Geneva, Switzerland), DNASTAR (DNASTAR Inc., University of California- San Francisco, San Francisco, CA) and BLAST version 2.1 (www.ncbi.nlm.nih.gov/BLAST/index.compat.html) computer programs.

Genomic DNA isolation

After the project was approved by the institutional review board and written informed consent had been obtained, peripheral blood samples were collected from 50 unrelated individuals without thyroid pathology. Genomic DNA was isolated from white blood cells by the SDS-proteinase K method.

Long PCR amplification and DNA sequencing

The long polymerase chain reaction (PCR) is suitable for amplification of long DNA templates. This approach was used to amplify the region containing intron 17/exon 18/intron 18/exon 19/intron 19 by primers situated in introns 17 and 19.

The primers used were 5′ gcagaggaaatcccaa 3′ (forward primer) and 5′ ctcagagaggctgcatagctt 3′ (reverse primer).

The long PCR were performed in 50 μL, using a standard elongase buffer (Invitrogen, Life Technologies, Carlsbad, CA), containing 150–200 ng of DNA, 1.5 mmol/L $MgSO_4$, 200 μmol/L of each dNTP, 1 μL Elongase Enzyme Mix (Invitrogen, Life Technologies), and 10 pmol of each forward and reverse primers.

Samples were heated to 94°C for 2 minutes, followed by 35 cycles of DNA denaturation (94°C for 30 seconds), annealing (55°C for 30 seconds), and polymerization (68°C for 7 minutes). Amplification was carried out in a MJ Research PTC 100 thermoblock (MJ Research, Watertown, MA). The amplified fragments were analyzed in a 1% agarose gel. The fragments for DNA sequencing were prepared with Concert gel purification system (Invitrogen, Life Technologies). The DNA sequencing was performed as those described above with the e18-F primer (Fig. 1).

Statistical analysis

A standard χ^2 analysis of the observed and expected genotypes was carried out in order to test for Hardy-Weinberg equilibrium (20).

Electronic database information

The nucleotide sequence data reported in this paper have been submitted to the GenBank database (www.ncbi.nlm.nih.gov/) under the accession numbers AF105683 and AY053519. Genomic sequence of the Tg gene: GenBank database accession numbers AF230666, AF230667, AF235100, AF305872 and NT_008150.

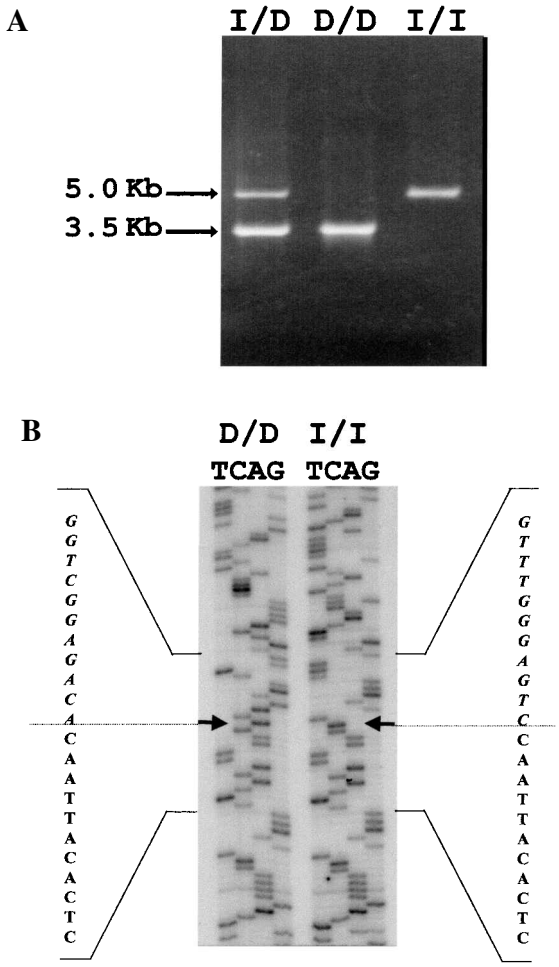

FIG. 2. **A**: Examples of the 1464-bp insertion/deletion polymorphic genotypes. Molecular weights are indicated in kilobases (kb). **B**: Partial nucleotide sequence of the 1464-bp insertion/deletion polymorphic region of the thyroglobulin gene, from two homozygous unrelated individuals. Sequence reactions were made with the e18-F primer. The arrows denote the position of insertion/deletion site. I, insertion; D, deletion.

Results

Analysis of the structural organization of the Tg gene has revealed an excellent correlation between the restriction physical map based on our λ clones, previously characterized (3,6–9), and the electronic map (DNASTAR program) from BAC sequences reported in GenBank except for the *Eco*RI fragment that contains exons 17 and 18. λ dash 62 (8) shows a band of 4.2 kb, whereas BAC shows a band of 2.7 kb. The length difference could be the result of either a polymorphic restriction site or to an insertion. We have previously sequenced the exon 18 and the first 277 nt of the intron 18 (8). Computer alignments demonstrated that 194 nt located inside intron 18 are not present in the BAC.

We extended our initial molecular studies to show the complete inclusion region by direct sequencing of λ dash 62 with several intronic primers. Sequence analysis showed an additional 1464-bp region located 83 bp downstream from exon 18 (Fig. 1). This segment is missing in the BAC and exhibits a high A+T content (62%) and a complicated sequence organization (Fig. 1). A 17-bp long motif (AAGAATTTTGGAGAACA) is found repeated twice throughout the sequence, located 791 and 849 bp downstream from exon 18. Three different types of 10-bp long palindromic sequences, ATTAGCTAAT, TTTTATAAAA and CAAATATTTG, were also found at positions 288, 870, and 1214, respectively. In addition, three short $(A)_n$ repeat traits along the sequence were identified. Comparison of the 1464-bp region with sequences deposited in the GenBank database, using Blast network service, failed to reveal significant levels of homology to any previously described sequences.

In order to study the possible polymorphism of these sequences we amplified them by long PCR using intronic primers. The genotyping was carried out in a population sample of unrelated Caucasian individuals (100 chromosomes). The amplification procedure generates two fragments of 3.5 and 5.0 kb that correspond to the exclusion or inclusion of the 1464-bp segment, respectively (Fig. 2a). The allele and genotype frequencies are summarized in Table 1. From the 50 samples analyzed, 18 were homozygous for the allele with the insertion, 12 were homozygous for the allele with the deletion, and 20 were heterozygous, giving allele frequencies of 0.56 (insertion) and 0.44 (deletion). The χ^2 analysis of observed and expected genotypes showed a nonsignificant p value, indicating that the sampled population does not deviate from Hardy-Weinberg equilibrium. Mendelian transmission of the alleles was verified in two families. The polymorphism was confirmed by sequence analysis of the 5.0- and 3.5-kb amplified fragments from homozygous samples (Fig. 2b).

TABLE 1. TEST OF HARDY-WEINBERG EQUILIBRIUM OF GENOTYPE FREQUENCIES OF THE 1464 BASE PAIR-INDEL POLYMORPHISM

	Observed		*Expected*	
Genotypes	*Number of subjects*	*Genotype frequency*	*Number of subjects*	*Genotype frequency*
I/I	18	0.36	15.7	0.314
I/D	20	0.40	24.6	0.492
D/D	12	0.24	9.7	0.194
χ^2 1.742	$p > 0.10$ (not significant)		$df = 1$	$n = 50$
	Allele frequencies: I: 0.56 D: 0.44.			

The expected genotype frequencies were calculated from the allele frequencies.
I, insertion; D, deletion.

Discussion

The Tg gene is a good example of a classic exon duplication (5) and presents structural characteristics that make it an interesting model to test theories about the potential role of introns. Introns are heterogeneous entities with different functional capacities and notable structural differences. In this context, we present here a detailed examination of a novel large Indel polymorphism of 1464 bp localized in intron 18 of the human Tg gene. This sequence contains high A+T content, two 17-bp long motif repeats, three different types of 10-bp long palindromic sequences as well as three short $(A)_n$ repeat traits (Fig. 1). The allele frequency showed a relatively high value of both alleles (Table 1).

The diallelic polymorphism of the DNA may be the result of a single nucleotide substitution, or the insertion and deletion of short stretches of sequences. However, in the Indel polymorphism it is impossible to distinguish if the variation is caused by an insertion in one sequence or a deletion in another. Large Indel polymorphisms have been rarely described (21). Genetic evidence indicates that the small additions and deletions can occur spontaneously during replication. Deletion and insertion also result from recombination events or activities of the transposable elements. Some members of the interspersed repetitive families in the human genome have been considered transposable elements such as Alu repeats (22), LINE1 family (23), and HERV sequences (24), which appear to have been created by retrotransposition (25). These elements can be transcribed into RNA and reverse-transcribed into cDNA, and subsequently, the cDNA can be reinserted into the genome at a new location A. GenBank database search found that the 1464-bp Indel polymorphism does not correspond to any known interspersed repetitive human sequence. More specifically, a detailed comparison between this 1464-bp sequence and full-length Alu, LINE1, and HERV sequences exhibits an nonsignificant degree of homology. Therefore, these results confirm that the new Indel polimorphism is not a transposable element. However, it is not possible to rule out that some ancient transposable element, not identified in the intron 18, might have been involved in the development of this polymorphism.

In conclusion, we report the identification of a novel member of the Indel polymorphism family. This new polymorphism would prove a convenient genetic marker to explore the 8q24 region of the human genome. Further studies in the field of molecular evolution and population genetics are necessary for elucidating the origin and mutational processes of this sequence.

Acknowledgments

C.M. Moya and F.M. Mendive are research fellows of the Universidad de Buenos Aires. C.M. Rivolta is a research fellow of the Argentine National Research Council (CONICET). H.M. Targovnik is a established investigator of the CONICET.

This work was supported by grants from Universidad de Buenos Aires (TB 80/98, B 087/2001), CONICET (0853/98), and FONCYT (05-00000-01591/ PICT 97, 05-08838/ PICT 2000/2001).

References

1. van de Graaf SAR, Ris-Stalpers C, Pauws E, Mendive FM, Targovnik HM, de Vijlder JJM 2001 Up-to-date with the human thyroglobulin sequence. J Endocrinol **170:**307–321.
2. van Ommen G-JB, Arnberg AC, Baas F, Brocas H, Sterk A, Tegelaers WHH, Vassart G, de Vijlder JJM 1983 The human thyroglobulin gene contains two 15–17 kb introns nears its 3′-end. Nucleic Acids Res **11:**2273–2285.
3. Targovnik HM, Pohl V, Christophe D, Cabrer B, Brocas H, Vassart G 1984 Structural organization of the 5′ region of the human thyroglobulin gene. Eur J Biochem **141:**271–277.
4. Baas F, van Ommen G-JB, Bikker H, Arnberg, AC, de Vijlder JJM 1986 The human thyroglobulin gene is over 300 kb long and contains introns of up to 64 Kb. Nucleic Acids Res **14:**5171–5186.
5. Parma J, Christophe D, Pohl V, Vassart G 1987 Structural organization of the 5′ region of the thyroglobulin gene. Evidence for intron loss and "exonization" during evolution. J Mol Biol **196:**769–779.
6. Targovnik H, Paz C, Corach D, Christophe D 1992 The 5′ region of the human thyroglobulin gene contains members of the Alu family. Thyroid **2:**321–324.
7. Mendive FM, Rivolta CM, Vassart G, Targovnik HM 1999 Genomic organization of the 3′ region of the human thyroglobulin gene. Thyroid **9:**903–912.
8. Moya CM, Mendive FM, Rivolta CM, Vassart G, Targovnik HM 2000 Genomic organization of the 5′ region of the human thyroglobulin gene. Eur J Endocrinol **143:**789–798.
9. Mendive FM, Rivolta CM, Moya CM, Vassart G, Targovnik HM 2001 Genomic organization of the human thyroglobulin gene. The complete intron-exon structure. Eur J Endocrinol **145:**485–496.
10. Ieiri T, Cochaux P, Targovnik HM, Suzuki M, Shimoda S-I, Perret J, Vassart G 1991 A 3′ splice site mutation in the thyroglobulin gene responsible for congenital goiter with hypothyroidism. J Clin Invest **88:**1901–1905.
11. Targovnik HM, Medeiros-Neto G, Varela V, Cochaux P, Wajchenberg BL, Vassart G 1993 A nonsense mutation causes human hereditary congenital goiter with preferential production of a 171-nucleotide-deleted thyroglobulin ribonucleic acid messenger. J Clin Endocrinol Metab **77:** 210–215.
12. Targovnik HM, Vono J, Billerbeck AEC, Cerrone GE, Varela V, Mendive F, Wajchenberg BL, Medeiros-Neto G 1995 A 138-nucleotide deletion in the thyroglobulin ribonucleic acid messenger in a congenital goiter with defective thyroglobulin synthesis. J Clin Endocrinol Metab **80:**3356–3360.
13. Targovnik HM, Frechtel GD, Mendive FM, Vono J, Cochaux P, Vassart G, Medeiros-Neto G 1998 Evidence for the segregation of three different mutated alleles of the thyroglobulin gene in a Brazilian family with congenital goiter and hypothyroidism. Thyroid **8:**291–297.
14. Hishinuma A, Takamatsu J, Ohyama Y, Yokozawa T, Kanno Y, Kuma K, Yoshida S, Matsuura N, Ieri T 1999 Two novel cysteine substitutions (C1263R and C1995S) of thyroglobulin cause a defect in intracellular transport of thyroglobulin in patients with congenital goiter and the variant type of adenomatous goiter. J Clin Endocrinol Metab **84:**1438–1444.
15. van de Graaf SAR, Ris-Stalpers C, Veenboer GJM, Cammenga M, Santos C, Targovnik HM, de Vijlder JJM, Medeiros-Neto G 1999 A premature stop codon in thyroglobulin mRNA results in familial goiter and moderate hypothyroidism. J Clin Endocrinol Metab **84:**2537–2542.
16. Targovnik HM, Rivolta CM, Mendive FM, Moya CM, Medeiros-Neto G 2001 Congenital goiter with hypothyroidism due to a 5′ splice site mutation in the thyroglobulin gene. Thyroid **11:**683–688.
17. Corral J, Martín C, Pérez R, Sánchez I, Mories MT, San Millan JL, Miralles JM, González-Sarmiento R 1993 Thyroglob-

ulin gene point mutation associated with non-endemic simple goitre. Lancet **341:**462–464.
18. Pérez-Centeno C, González-Sarmiento R, Mories MT, Corrales JJ, Miralles-García JM 1996 Thyroglobulin exon 10 gene point mutation in a patient with endemic goiter. Thyroid **6:**423–427.
19. Gonzalez-Sarmiento R, Corral J, Mories MT, Corrales JJ, Miguel-Velado E, Miralles-Garcia JM 2001 Monoallelic deletion in the 5′ region of the thyroglobulin gene as a cause of sporadic nonendemic simple goiter. Thyroid **11:**789–793.
20. Sokal RR, Rohlf FJ 1995 Biometry: The Principles and Practice of Statistics in Biological Research. W.H. Freeman, New York.
21. Rigat B, Hubert C, Alhenc-Gelas F, Cambien F, Corvol P, Soubrier F 1990 An insertion/deletion polymorphism in the angiotensin I-converting enzyme gene accounting for half the variance of serum enzyme levels. J Clin Invest **86:**1343–1346.
22. Schmid CW 1996 Alu: Structure, origin, evolution, significance and function of one-tenth of human DNA. Prog Nucleic Acid Res Mol Biol **53:**283–319.
23. Dombroski BA, Scott AF, Kazazian HH 1993 Two additional potential retrotransposons isolated from a human L1 subfamily that contains an active retrotransposable element. Proc Natl Acad Sci USA **90:**6513–6517.
24. Feuchter-Murthy AE, Freeman JD, Mager DL 1993 Splicing of a human endogenous retrovirus to a novel phospholipase A2 related gene. Nucleic Acids Res **21:**135–143.
25. Sassaman DM, Dombroski BA, Moran JV, Kimberland ML, Naas TP, De Berardinis RJ, Gabriel A, Swergold GD, Kazazian HH 1997 Many human L1 elements are capable of retrotransposition. Nat Genet **16:**37–43.

Address reprint requests to:
Dr. Héctor M. Targovnik
Laboratorio de Biología Molecular
Cátedra de Genética y Biología Molecular
Facultad de Farmacia y Bioquímica
Universidad de Buenos Aires
Av. Córdoba 2351
4to piso-sala 5
1120-Buenos Aires
Argentina

E-mail: htargovn@huemul.ffyb.uba.ar