GTF2IRD2 is located in the Williams–Beuren syndrome critical region 7q11.23 and encodes a protein with two TFII-I-like helix–loop–helix repeats

Aleksandr V. Makeyev*, Lkhamsuren Erdenechimeg†, Ognoon Mungunsukh†, Jutta J. Roth†‡, Badam Enkhmandakh†, Frank H. Ruddle†, and Dashzeveg Bayarsaihan§¶

†Department of Molecular, Cellular, and Developmental Biology, Yale University, 266 Whitney Avenue, New Haven, CT 06520; *Department of Genetics and Development, Columbia University, 701 West 168th Street, NY 10032; ‡Department of Genetics and General Biology, University of Salzburg, Hellbrunnerstrasse 34, 5020 Salzburg, Austria; and §Department of Molecular, Cellular, and Craniofacial Biology, Birth Defect Center, University of Louisville, 501 South Preston Street, Louisville, KY 40202

Contributed by Frank H. Ruddle, June 10, 2004

Williams–Beuren syndrome (also known as Williams syndrome) is caused by a deletion of a 1.55- to 1.84-megabase region from chromosome band 7q11.23. *GTF2IRD1* and *GTF2I*, located within this critical region, encode proteins of the TFII-I family with multiple helix–loop–helix domains known as I repeats. In the present work, we characterize a third member, *GTF2IRD2*, which has sequence and structural similarity to the *GTF2I* and *GTF2IRD1* paralogs. The ORF encodes a protein with several features characteristic of regulatory factors, including two I repeats, two leucine zippers, and a single Cys-2/His-2 zinc finger. The genomic organization of human, baboon, rat, and mouse genes is well conserved. Our exon-by-exon comparison has revealed that *GTF2IRD2* is more closely related to *GTF2I* than to *GTF2IRD1* and apparently is derived from the *GTF2I* sequence. The comparison of *GTF2I* and *GTF2IRD2* genes revealed two distinct regions of homology, indicating that the helix–loop–helix domain structure of the *GTF2IRD2* gene has been generated by two independent genomic duplications. We speculate that *GTF2I* is derived from *GTF2IRD1* as a result of local duplication and the further evolution of its structure was associated with its functional specialization. Comparison of genomic sequences surrounding *GTF2IRD2* genes in mice and humans allows refinement of the centromeric breakpoint position of the primate-specific inversion within the Williams–Beuren syndrome critical region.

Williams–Beuren syndrome (WBS, also known as Williams syndrome) is a neurodevelopmental disorder caused by a 1.55- to 1.84-megabase deletion at 7q11.23. Patients carrying this disorder exhibit supravalvular and aortic stenosis, growth retardation, premature aging of the skin, mental retardation, and dental malformations (1–4). Several loci exist within the deleted region that encode transcription factors and chromatin-remodeling proteins (1, 2). Two such genes, *GTF2IRD1* and *GTF2I*, encode proteins belonging to the TFII-I family of transcription factors, characterized by the presence of multiple helix–loop–helix (HLH) domains known as I repeats (5–11). Both paralogs are highly conserved in vertebrates and have a broad expression pattern in adult and embryonic tissues (12, 13). Protein products of these genes are implicated in gene regulation through interactions with different tissue-specific transcription factors and chromatin-remodeling complexes (5, 14).

Discovery of the *GTF2IRD2* gene, an additional member of the TFII-I family, and its pseudogenes has been reported recently (1, 2, 15). Here, we describe the genomic structural organization of human and mouse *GTF2IRD2* orthologs.

Materials and Methods

Nucleotide sequence databases were searched by using standard nucleotide–nucleotide BLAST and MEGABLAST with standard parameters at the National Center for Biotechnology Information BLAST Server (www.ncbi.nlm.nih.gov/BLAST). The entire coding region of the mouse Gtf2ird2 was sequenced from the mouse expressed sequence tag (EST) clone IMAGE 5033059 (GenBank accession no. BI156030). Genomic sequences were analyzed with the latest version of REPEATMASKER (A. F. A. Smit and P. Green, www.repeatmasker.org). Promoter analysis was performed with PROMOTERINSPECTOR (16) and MATINSPECTOR RELEASE 7.2 (17). The protein parameters were analyzed by using the PROTPARAM, CLUSTALW, PROSITE, and PREDICT PROTEIN programs.

Results and Discussion

Sequence Analysis of *GTF2IRD2* Genes. The dbEST public database of the National Center for Biotechnology Information was searched for human and mouse ESTs bearing sequence similarity to the mouse *Gtf2i* sequence (GenBank accession no. AY030291). Several cDNA entries were identified, including the mouse IMAGE 5033059 and 5043754 clones (GenBank accession nos. BI156030 and BI103332; UniGene Cluster Mm.218744) and human clones IMAGE 3310920, 89677, and 2710422 (GenBank accession nos. BF001292, AA376914, and AW015648; UniGene Cluster Hs.399978). These clones were completely sequenced and shown to be identical with the previously reported sequences of mouse *Gtf2ird2* (GenBank accession no. AY014963) and human *GTF2IRD2* (GenBank accession no. BC047706; RefSeq: NM_173537) genes.

The assembled *Gtf2ird2* and *GTF2IRD2* cDNAs contain ORFs that encode putative proteins of 936 aa in mouse and 949 aa in human with a calculated molecular mass of 105 kDa and pI 5.83 or 107 kDa and pI of 5.53, respectively. Almost 79% identity and >90% similarity exist between human and mouse proteins (data not shown). The ORF encodes a protein with several features characteristic of regulatory factors, including two TFII-I-like HLH domains (amino acids 107–182 and 333–407 in human and 104–180 and 329–403 in mouse sequence), two leucine zippers (amino acids 23–44 and 776–798 in human and 21–42 and 750–792 in mouse sequence), and a single Cys-2/His-2 zinc finger (amino acids 435–471 and 431–467 in human and mouse protein, respectively) (Fig. 1*A*). The presence of these domains suggests that GTF2IRD2 possesses complex protein-binding properties.

To elucidate the genomic structure of *GTF2IRD2*, the complete human and mouse cDNA sequences were compared with publicly available genomic sequences from the two human bacterial artificial chromosome clones RP11-813J7 and CTA-

Abbreviations: HLH, helix–loop–helix; WBS, Williams–Beuren syndrome.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. AY260739, AY116023, BK005162, and BK005163).

¶To whom correspondence should be addressed. E-mail: bayarsaihan@yale.edu.

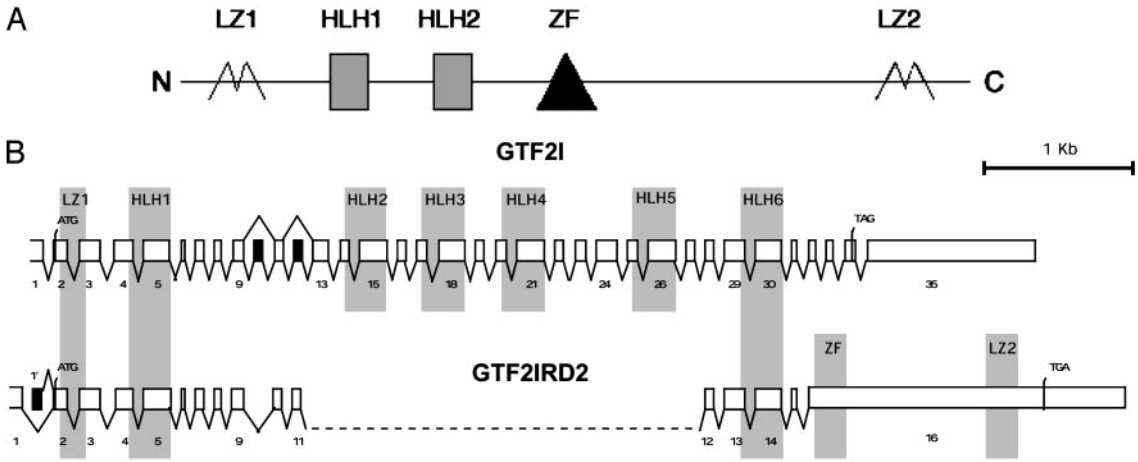

Fig. 1. Protein domains and genomic structure of GTF2IRD2. (*A*) Structural organization of the GTF2IRD2 protein. HLH, HLH repeats; LZ, leucine zipper; ZF, zinc finger domain. (*B*) Homology of *GTF2IRD2* and *GTF2I* genes. Comparison of exon–intron structure. Exons are shown as boxes; the black exons represent alternatively spliced sequences of the transcript. The major splicing patterns are indicated by the exon–exon connections shown below each of the diagrams, and the alternative splicing combinations are indicated above the diagrams. The regions of the transcript encoding the HLH domains are included in the shaded box. The 5′ termini of all cDNAs begin at the 5′-most extent of the EST sequences (left is open). The exons are drawn to scale, and the introns, which vary greatly in size, are represented in a uniform manner. (Scale bar = 1 kb.)

350L10 (GenBank accession nos. AC083884 and AC005098) and from the mouse clone RP23-15K13 (GenBank accession no. AC093346), respectively. The sequence of *GTF2IRD2* comprises 16 exons extending over 57 kb. The murine ortholog *Gtf2ird2* has a similar exon–intron structure, although it is more compact and spans ≈34 kb. Exon 16 contains the translational stop codon and the 3′-UTR (517 bp in humans and 432 bp in mice) with a poly(A) signal (ATTAAA in humans and AATAAA in mice).

GTF2IRD2 has significant structural similarity to other *GTF2I* family members. The relationship of *GTF2IRD2* to *GTF2I* is obvious from comparison of both exon–intron structure (Fig. 1*B*) and sequence similarity (Fig. 2) of these two genes. Genomic structural analysis revealed two regions of homology: (*i*) exons 2–11 of *GTF2IRD2* correspond to exons 2–12 of *GTF2I* (the optional exon 10 of *GTF2I* is absent in the genomic sequence of *GTF2IRD2*) and (*ii*) exons 12–15 of *GTF2IRD2* correspond to exons 28–31 of *GTF2I* (Fig. 1*B*). Corresponding exons demonstrate a high level of sequence similarity (75–91% identity and 89–96% similarity), the same phase and identical length (with the single exception of 5′ intron–exon boundary sliding in exon 4 of *GTF2IRD2*) (Fig. 2). The sequence around the first methionine (GAACAATGG) in *GTF2IRD2* is in agreement with the Kozak consensus and corresponds exactly to the translational start in *GTF2I*. The remarkable conservation of sequence and exon–intron architecture of two long segments of the *GTF2IRD2* and *GTF2I* paralogs strongly support their common origin.

Several translated nucleotide sequences in the public database share high similarity to the GTF2IRD2. Over 35% identity and 53% similarity occurs with the GenBank entry EAA09584 in a 539-aa overlap, 29% identity and 47% similarity with the entry AAO21376 in a 457-aa overlap, and 71% identity and 88% similarity with the entry AAG15589 in a 65-aa overlap (data not shown). In addition, several proteins of unknown function show a weak similarity, including EbiP438 from *Anopheles gambiae* (GenBank accession no. AAA09584.1; 33% identity), human KIAA0766 (GenBank accession no. BAA34486.1; 27% identity), human KIAA1353 (GenBank accession no. BAA92591.1; 26% identity), and human transposase-like protein (GenBank accession no. AF205600.1; 26% identity) (data not shown).

Based on the existence of highly homologous sequences (>80% identity at the protein level) from cow, trout, chicken, and pig, we concluded that *GTF2IRD2* is well conserved within the vertebrate lineage. In fact, we were able to find baboon and rat genomic sequences that included predicted cDNAs with high homology to the human and mouse genes.

Chromosomal Location of *GTF2IRD2* Orthologs. Chromosomal instability at 7q11.23 results from its complex genomic structure; the region contains three large segmental duplications (centromeric, medial, and telomeric) and each of them is composed of three different blocks (A, B, and C) (18). During chromosome pairing in cell divisions, these duplicated segments may favor unequal crossing-over or nonallelic homologous recombination, causing deletions or paracentric inversions (15). The duplications are also present in non-human primates, including chimpanzees, gorillas, orangutans, and gibbons, but are absent in mice (18, 19). Analysis of the murine locus on chromosome 5G1 (clone RP23–15K13, accession no. AC093346) revealed that *Gtf2ird2* and *Gtf2i* are ≈19.7 kb apart, whereas human *GTF2IRD2* and *GTF2I* are separated by 35 kb. Both paralogs are separated by the *Ncf1* locus (Fig. 3). *Gtf2ird2* is arranged in an opposite orientation to *Gtf2ird1* and *Gtf2I,* respectively.

Whereas the mouse genome contains only one *Gtf2ird2*, three *GTF2IRD2* loci are contained in the human 7q11.23 region, which is syntenic to mouse 5G1 (15). In addition to the functional *GTF2IRD2* gene, the centromeric (B_c) and telomeric (B_t) repeats contain putative pseudogenes *GTF2IRD2P1* and *GTF2IRD2P2*, respectively (Fig. 3). Despite very high sequence homology, single nucleotide substitutions in exon 16 would allow distinction of mRNA products of all three loci. However, it is unlikely that *GTF2IRD2P1* is transcribed because of the deletion of exons 1 and 2, whereas the transcription status of *GTF2IRD2P2* is unknown. Our promoter analysis shows that a 9-bp deletion found in the upstream genomic sequence does not affect the promoter region of *GTF2IRD2P2* (see figure 3*A* in ref. 15).

Comparison of genomic sequences surrounding *GTF2IRD2* genes in mice and humans allows refinement of the position of centromeric breakpoint of the primate-specific inversion of the WBS critical region. In the work of Valero *et al.* (18), the boundary of synteny has been localized between *Wbscr16* (GenBank accession no. AA008727) and *Wbscr17* (GenBank accession no. AA388221) in the mouse. However, a putative human ortholog of the next mouse gene, *Gats*, can be found far beyond the B_t repeat (Fig. 3). This extension of homology moves the

breakpoint into the poorly characterized region between the *Gats* and *Wbscr17* genes. Furthermore, the human *WBSCR16* gene has an inverted orientation with respect to other genes of the syntenic group (Fig. 3). This indicates that the WBS chromosomal segment was subjected to a more complicated reorganization during primate evolution, including not only the inversion of the whole region and insertion of low-copy-number repeats, but also local rearrangements.

Promoter Analysis of *GTF2IRD2* Genes. The 5′-most extent of the EST sequences in UniGene clusters Hs.399978 and Mm.218744 indicates that mouse transcripts have a longer 5′-UTR than human transcripts as a result of 5′ extension of the first noncoding exon. The transcription start sites of human and mouse mRNAs predicted by PROMOTERINSPECTOR are 194 and 237 bp upstream of the translational start sites, respectively. No known promoter motifs, such as TATA boxes, initiators, or downstream

A

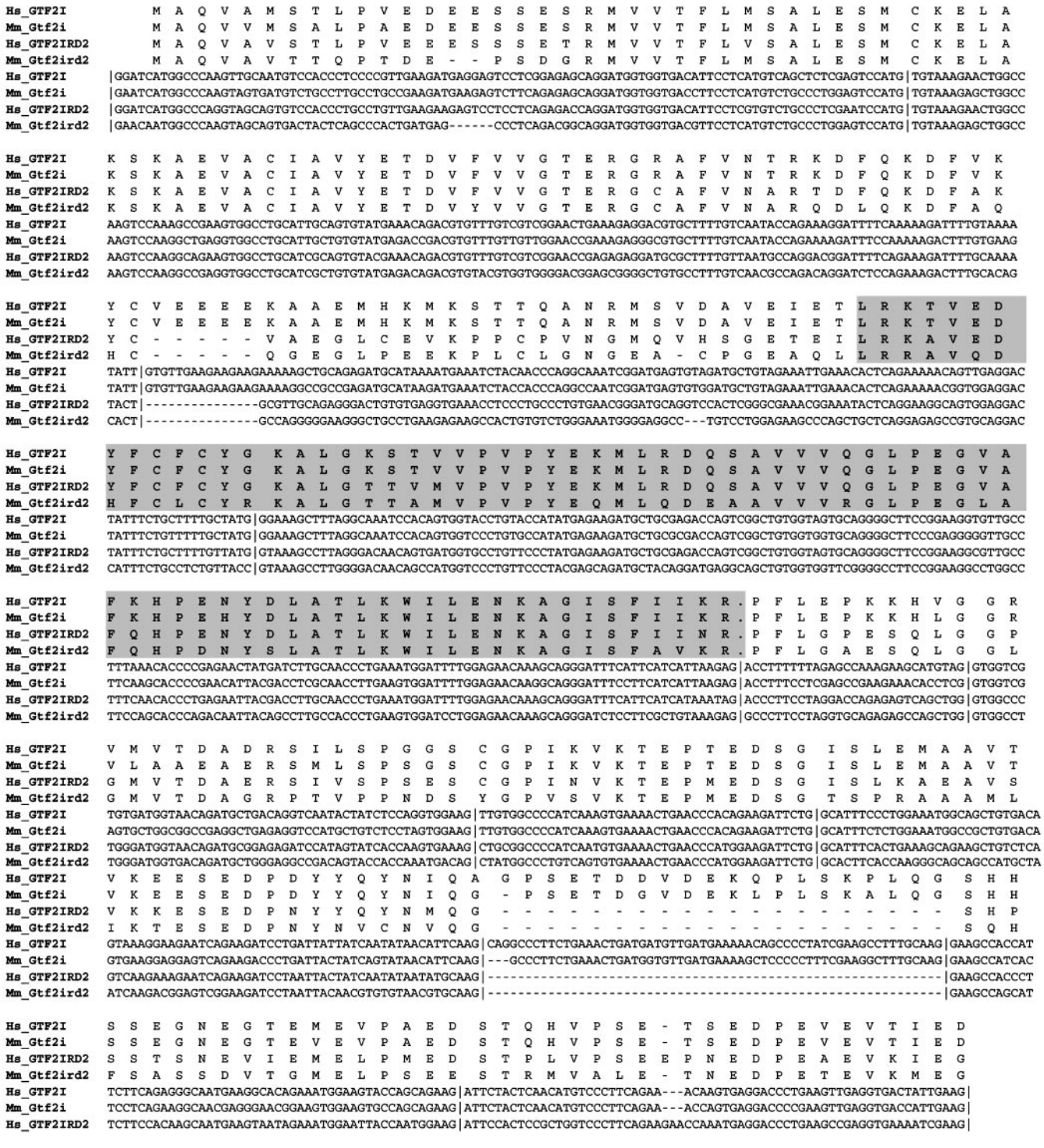

Fig. 2. (*Figure continues on the opposite page.*)

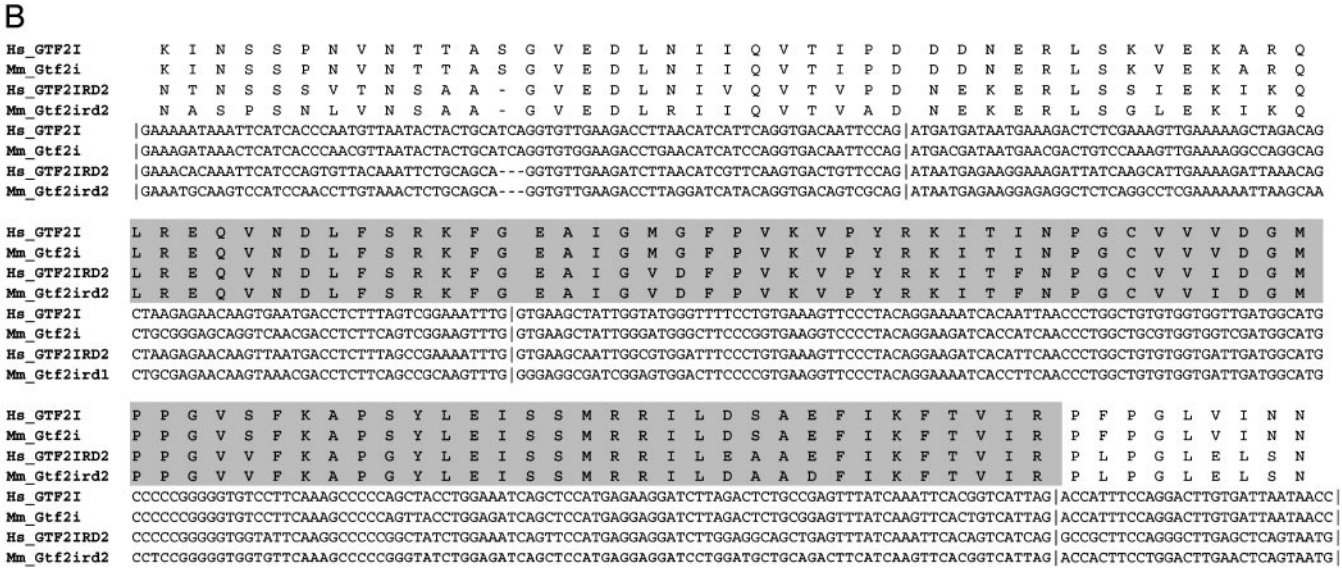

Fig. 2. Two regions of sequence homology between *GTF2IRD2* and *GTF2I* genes. (*A*) Sequence alignment of the N-terminal region of conservation (exons 2–11 of *GTF2IRD2* and *Gtf2ird2* and exons 2–12 of *GTF2I* and *Gtf2i*). (*B*) Internal region of conservation (exons 12–15 of *GTF2IRD2* and *Gtf2ird2* and exons 28–31 of *GTF2I* and *Gtf2i*). TFII-I-like HLH repeats (I repeats) are shown in gray background. Hs, *Homo sapiens*; Mm, *Mus musculus*.

promoter elements, exist in the vicinity of mouse or human transcription start sites. However, a putative TFIIB recognition element was found in positions −63 to −58 in human and −62 to −57 in mouse genomic sequences (Fig. 4*A*). These motifs are also present in corresponding positions in baboon and rat genomic sequences, and it could be that these elements determine the difference in position of the transcription start site between primates and rodents. Binding sites for GC and CCAAT box-binding proteins were identified in human (−229 to −215, −197 to −187, and +22 to +36) and in mouse (−198 to −184), but they are not conserved between species. In contrast, MATINSPECTOR identified specific response elements organized as promoter modules that are conservative in human, baboon, mouse, and rat sequences despite low (28%) overall sequence homology between primate and rodent sequences (Fig. 4*A*).

An alternative *GTF2IRD2* transcript has been found in a cDNA library prepared from normal lung epithelial cells (GenBank accession no. BM973984). This transcript includes a proximal optional exon that is spliced normally with exon 2 (Fig. 1, exon 1′). Although such transcripts were not found in mouse and

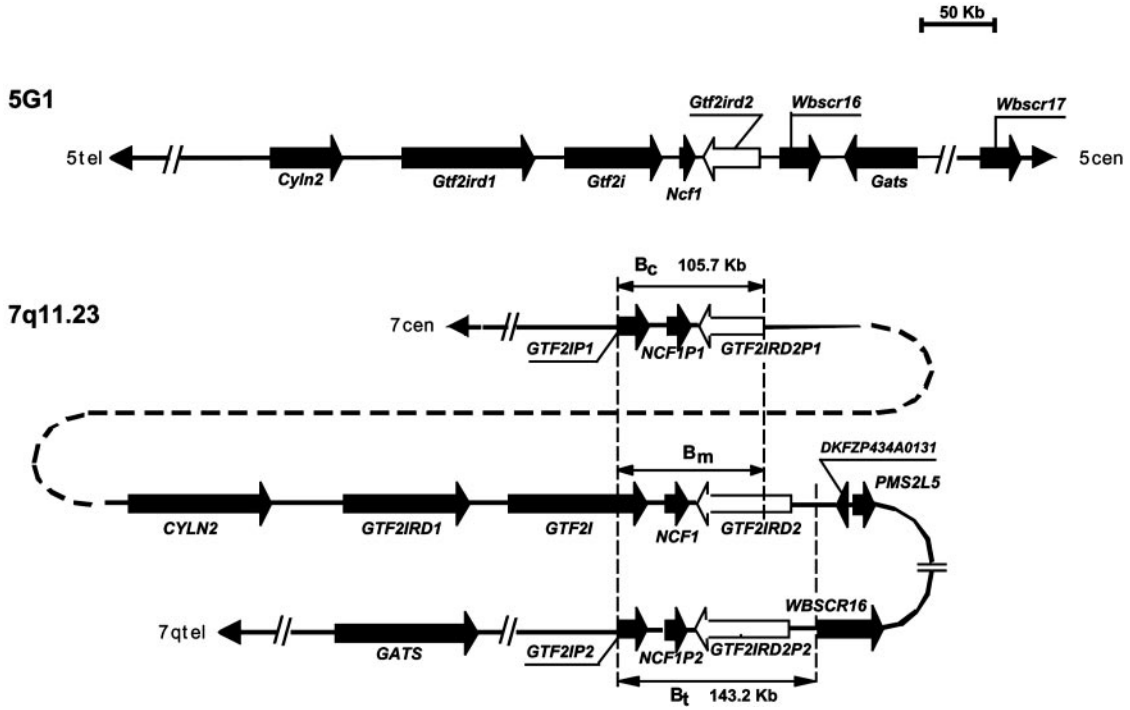

Fig. 3. Comparative representation of the genomic region surrounding the *GTF2IRD2* loci on human chromosome 7q11.23 and in the syntenic region of mouse chromosome 5G1. *GTF2IRD2* loci are shown as white block arrows (5′–3′ direction), whereas all other genes are shown as black block arrows.

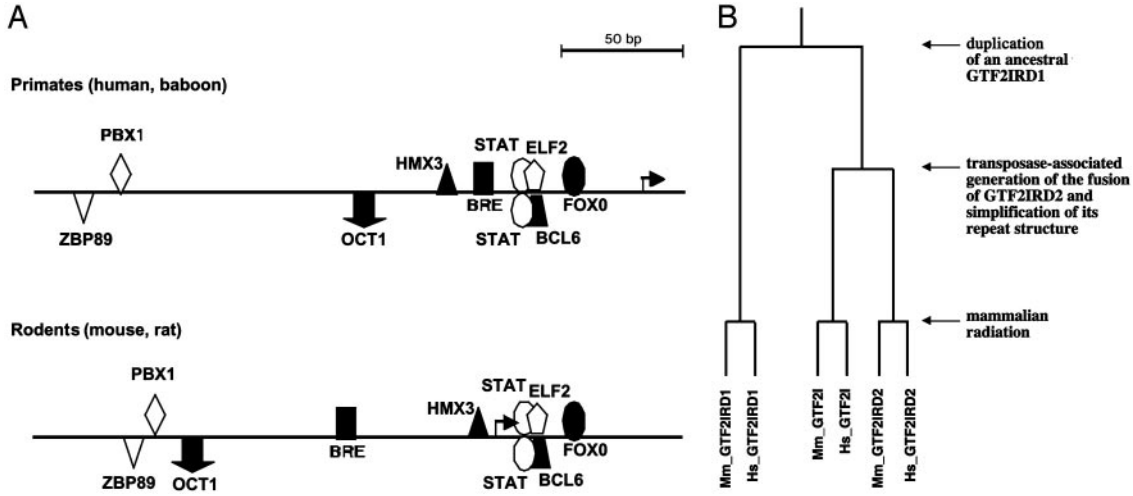

Fig. 4. Promoter organization of *GTF2IRD2* genes and their evolutionary relationship with other members of the GTF2I family. (*A*) Regulatory elements that can contribute to basal and tissue-specific transcription of primate and rodent *GTF2IRD2* genes. The diagram shows the average relative distances calculated either for human and baboon or mouse and rat sequences. Arrow indicates transcription start site. BCL6, POZ/zinc finger protein, transcriptional repressor; BRE, TFIIB recognition element; ELF2, Ets-family member ELF-2/NERF1a; FOXO, Fkh domain factor FKHRL1; HMX3, H6 homeodomain HMX3/Nkx5.1; OCT1, octamer-binding factor 1; PBX1, homeodomain factor Pbx1; STAT, signal transducers and activators of transcription; ZBP89, zinc finger transcription factor ZBP-89. (*B*) Hypothetic phylogenetic tree of the *GTF2I*-related genes. The evolutionary distances between sequences are not drawn to scale.

rat dbESTs, an almost identical sequence was identified within baboon intron 1 (data not shown), indicating the possible existence of a primate-specific alternative mRNA variant.

Evolution of the TFII-I Family. Our analysis clearly indicates that HLH repeats 1 and 2 of GTF2IRD2 are homologous to the HLH1 repeats 1 and 6 of TFII-I/GTF2I, respectively. We have shown that the six HLH repeats of GTF2IRD1 and GTF2I had a different duplication history (20). We have also identified a partial sequence of a *GTF2I*-related gene containing five HLH repeats in *Danio rerio* and *Takifugu rubripes* (data not shown). This sole fish sequence is very similar to *GTF2IRD1*, which is likely to represent the oldest member of the *GTF2I* family (Fig. 4*B*). Consequently, we speculate that *GTF2I* is derived from *GTF2IRD1* as a result of local duplication, and the further evolution of its structure was associated with its functional specialization. Our exon-by-exon comparison has revealed that *GTF2IRD2* is more closely related to *GTF2I* than to *GTF2IRD1* (Figs. 1*B* and 2) and apparently is derived from the *GTF2I* sequence. The origin of the *GTF2IRD2* gene and its opposite genomic orientation are not clear at present, but an unusual C-terminal CHARLIE8-like domain that is absent in other members of TFII-I family suggests that its transposase activity has generated a functional fusion gene (21). The acquired C-terminal domain probably provides some new functions to the GTF2IRD2 protein that do not require multiple HLH repeats and therefore made possible the loss of the four central repeats as a result of structural simplification. We speculate that the formation of *GTF2IRD2* was finished before mammalian radiation (Fig. 4*B*).

In this study we report the genomic organization of *GTF2IRD2,* which encodes a protein with structural similarity to the N-terminal end of TFII-I. We have also identified mouse, rat, and baboon orthologs that share significant similarity with the human sequence. The order and orientation of *GTF2IRD1*, *GTF2I*, and *GTF2IRD2* is conserved between human and mouse. Structurally, members of TFII-I family possess multiple HLH repeat domains and a leucine zipper motif. Recent data indicate that they are implicated in gene regulation through interactions with tissue-specific transcription factors and chromatin-remodeling complexes. TFII-I factors physically and functionally interact with PIASxβ and HDAC3, suggesting a complex interplay between TFII-I family members and histone modification and SUMOylation (22–24). GTF2I/TFII-I forms a complex with HDAC1, HDAC2, and BHC110 and is involved in transcriptional repression (14). GTF2IRD1/BEN was proposed to play an important role in fiber-specific muscle gene expression as a repressor involving MEF2C and NcoR (25). It also interacts with the retinoblastoma protein (Rb), an important regulator of cell cycle and development (26). We have shown that GTF2IRD1 represses transcriptional activity of TFII-I by a two-step competition mechanism involving a cytoplasmic shuttling factor and a nuclear cofactor required for transcriptional activation of GTF2I (27). Recent work indicates dynamic spatial and temporal expression patterns of the members of TFII-I family throughout embryonic development of the mouse (12, 13).

The *GTF2IRD2* locus is retained in the common 1.55-megabase deletion, but it is deleted in WBS patients with the rarer 1.84-megabase deletions (15). Therefore, in this longer deletion, all three loci of the TFII-I family become haploid and the lack of the *GTF2IRD2* allele could contribute to the WBS phenotype. Recent analysis suggests that *GTF2IRD2* and *GTF2I* contribute to deficits in visual spatial functioning (28). Other studies implicate *GTF2I* in the mental retardation of WBS (29).

Note. While this work was in preparation, we learned that the *GTF2IRD2* gene analysis was reported by Tipney *et al.* (21).

We thank Drs. Dmitry Nurminsky and Nyam-Osor Chimge for critical reading of the manuscript. J.J.R. is a fellow of the Doctoral Scholarship Program of the Austrian Academy of Sciences.

1. Perez Jurado, L. A. (2003) *Horm. Res.* **59,** 106–113.
2. Tassabehji, M. (2003) *Hum. Mol. Genet.* **12,** R229–R237.
3. Korenberg, J. R., Chen, X. N., Hirota, H., Lai, Z., Bellugi, U., Burian, D., Roe, B. & Matsuoka, R. (2000) *J. Cogn. Neurosci.* **12,** 89–107.
4. Mervis, C. B. (2003) *Dev. Neuropsychol.* **23,** 1–12.
5. Roy, A. L. (2001) *Gene* **274,** 1–13.
6. Wang, Y. K., Perez Jurado, L. A. & Francke, U. (1998) *Genomics* **48,** 163–170.

7. Perez Jurado, L. A., Wang, Y. K., Peoples, R., Coloma, A., Croces, J. & Francke, U. (1998) *Hum. Mol. Genet.* **7,** 325–334.
8. Osborne, L. R., Campbell, T., Daradich, A., Scherer, S. W. & Tsui, L. C. (1999) *Genomics* **57,** 279–284.
9. Franke, Y., Peoples, R. J. & Francke, U. (1999) *Cytogenet. Cell Genet.* **86,** 296–304.
10. Tassabehji, M., Carette, M., Wilmot, C., Donnai, D., Read, A. P. & Metcalfe, K. (1999) *Eur. J. Hum. Genet.* **7,** 737–747.
11. Bayarsaihan, D. & Ruddle, F. H. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 7342–7347.
12. Bayarsaihan, D., Bitchevaia, N., Enkhmandakh, B., Tussie-Luna, M. I., Leckman, J. F., Roy, A. & Ruddle, F. H. (2003) *Gene Expr. Patterns* **3,** 137–143.
13. Enkhmandakh, B., Bitchevaia, N., Ruddle, F. H. & Bayarsaihan, D. (2004) *Gene Expr. Patterns* **4,** 25–28.
14. Hakimi, M. A., Dong, Y., Lane, W. S., Speicher, D. W. & Shiekhattar, R. (2003) *J. Biol. Chem.* **278,** 7234–7239.
15. Bayes, M., Magano, L. F., Rivera, N., Flores, R. & Perez Jurado, L. A. (2003) *Am. J. Hum. Genet.* **73,** 131–151.
16. Scherf, M., Klingenhoff, A. & Werner, T. (2000) *J. Mol. Biol.* **297,** 599–606.
17. Quandt, K., Frech, K., Karas, H., Wingender, E. & Werner, T. (1995) *Nucleic Acids Res.* **23,** 4878–4884.
18. Valero, M. C., de Luis, O., Cruces, J. & Perez Jurado, L. A. (2000) *Genomics* **69,** 1–13.
19. DeSilva, U., Massa, H., Trask, B. J. & Green, E. D. (1999) *Genome Res.* **9,** 428–436.
20. Bayarsaihan, D., Dunai, J., Greally, J. M., Kawasaki, K., Sumiyama, K., Enkhmandakh, B., Shimizu, N. & Ruddle, F. H. (2002) *Genomics* **79,** 137–143.
21. Tipney, H. J., Hinsley, T. A., Brass, A., Metcalfe, K., Donnai, D. & Tassabehji, M. (2004) *Eur. J. Hum. Genet.* **12,** 551–560.
22. Tussie-Luna, M. I., Bayarsaihan, D., Seto, E., Ruddle, F. H. & Roy, A. L. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 12807–12812.
23. Tussie-Luna, M. I., Michel, B., Hakre, S. & Roy, A. L. (2002) *J. Biol. Chem.* **277,** 43185–43193.
24. Wen, Y. D., Cress, W. D., Roy, A. L. & Seto, E. (2003) *J. Biol. Chem.* **278,** 1841–1847.
25. Polly, P., Haddadi, L. M., Issa, L. L., Subramaniam, N., Palmer, S. J., Tay, E. S. & Hardeman, E. C. (2003) *J. Biol. Chem.* **278,** 36603–36610.
26. Yan, X., Zhao, X., Qian, M., Guo, N., Gong, X. & Zhu, X. (2000) *Biochem. J.* **345,** 749–757.
27. Tussie-Luna, M. I., Bayarsaihan, D., Ruddle, F. H. & Roy, A. L. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 7789–7794.
28. Hirota, H., Matsuoka, R., Chen, X. N., Salandanan, L. S., Lincoln, A., Rose, F. E., Sunahara, M., Osawa, M., Bellugi, U. & Korenberg, J. R. (2003) *Genet. Med.* **5,** 311–321.
29. Morris, C. A., Mervis, C. B., Hobart, H. H., Gregg, R. G., Bertrand, J., Ensing, G. J., Sommer, A., Moore, C. A., Hopkin, R. J., Spallone, P. A., *et al.* (2003) *Am. J. Med. Genet.* **123,** 45–59.