Comparison of *TFII-I* gene family members deleted in Williams-Beuren syndrome

TIMOTHY A. HINSLEY,[1,2] PAMELA CUNLIFFE,[1] HANNAH J. TIPNEY,[1] ANDREW BRASS,[2,3] AND MAY TASSABEHJI[1]

[1]Academic Department of Medical Genetics, St. Mary's Hospital, Manchester M13 OJH, United Kingdom
[2]School of Biological Science and [3]Department of Computer Science, University of Manchester, Manchester M13 9PL United Kingdom

Abstract

Williams-Beuren syndrome (WBS) is a neurological disorder resulting from a microdeletion, typically 1.5 megabases in size, at 7q11.23. Atypical patients implicate genes at the telomeric end of this multigene deletion as the main candidates for the pathology of WBS in particular the unequal cognitive profile associated with the condition. We recently identified a gene (*GTF2IRD2*) that shares homology with other members of a unique family of transcription factors (TFII-I family), which reside in the critical telomeric region. Using bioinformatics tools this study focuses on the detailed assessment of this gene family, concentrating on their characteristic structural components such as the leucine zipper (LZ) and I-repeat elements, in an attempt to identify features that could aid functional predictions. Phylogenetic analysis identified distinct I-repeat clades shared between family members. Linking functional data to one such clade has implicated them in DNA binding. The identification of PEST, synergy control motifs, and sumoylation sites common to all family members suggest a shared mechanism regulating the stability and transcriptional activity of these factors. In addition, the identification/isolation of short truncated isoforms for each TFII-I family member implies a mode of self-regulation. The exceptionally high identity shared between GTF2I and GTF2IRD2, suggests that heterodimers as well as homodimers are possible, and indicates overlapping functions between their respective short isoforms. Such cross-reactivity between GTF2I and GTF2IRD2 short isoforms might have been the evolutionary driving force for the 7q11.23 chromosomal rearrangement not present in the syntenic region in mice.

Keywords: Williams-Beuren Syndrome; GTF2I; GTF2IRD1; GTF2IRD2; short isoforms; PEST sequence; sumoylation sites; synergy control motif

Williams-Beuren syndrome (WBS, MIM 194050) is a neurological disorder associated with physical, behavioral, and cognitive abnormalities (Williams et al. 1961; Beuren et al. 1962). The complex phenotype includes an uninhibited friendly personality, mental retardation (overall IQ in the 50–60 range), and an unequal cognitive profile (WBSCP) where verbal tasks outstrip spatial tasks (Bennett et al. 1978; Udwin and Yule 1991; Bellugi et al. 2000; Donnai and Karmiloff-Smith 2000; Jones et al. 2000). Physical abnormalities include a characteristic dysmorphic face, retarded growth, hypercalcemia, hyperacusis, premature aging, and congenital heart defects of which supravalvular aortic stenosis (SVAS) is the most common.

WBS occurs mainly sporadically, with a frequency of ~1/20,000 live births (Morris et al. 1988), and is caused by a microdeletion at chromosome 7q11.23 (Ewart et al. 1993b). The deleted region is unstable due to the presence of flanking low copy repeats (LCRs) that share high sequence identity (>97%); one centromeric, one medial, and a telomeric LCR (Perez Jurado et al. 1996, 1998; Robinson et al. 1996; Osborne et al. 1997). The proposed mechanism for the chromosomal abnormality in WBS is thought to be un-

equal recombination between these nonsyntenic LCRs during meiosis (Dutly and Schinzel 1996; Perez Jurado et al. 1996; Baumer et al. 1998; Lupski 1998; Stankiewicz and Lupski 2002). This explains the sporadic nature of the disorder and the homogeneous size of the deletion (~1.5 Mb) detected in most patients. Inversion rearrangements in parents of individuals with WBS have also been reported at this locus and probably arise through the same mechanism (Osborne et al. 2001).

WBS was mapped to 7q11.23 after a SVAS patient was identified with a translocation that disrupted the elastin (*ELN*) gene at this locus (Curran et al. 1993; Ewart et al. 1993a). It has since been defined as a contiguous gene deletion disorder with *ELN* hemizygosity causing SVAS and other deleted genes resulting in the additional phenotypes (Ewart et al. 1993b). The identification of genes within the WBS region is ongoing but, to date, only *ELN* hemizygosity has been unambiguously shown to result in a WBS phenotype (Francke 1999; Osborne 1999). In genotype–phenotype correlations, rare patients with atypical deletions displaying only subsets of the WBS phenotype have implicated genes residing in the telomeric region of the common deletion in other features of WBS, including the cognitive and behavioral phenotypes (Botta et al. 1999; Tassabehji et al. 1999; Korenberg et al. 2000; Gagliardi et al. 2003; Hirota et al. 2003). This region contains members of a novel family of transcription factors (TFII-I family) containing both a putative leucine zipper (LZ) and unique I-repeat motifs (Fig. 1A). Members of the TFII-I family all cluster at the telomeric end of the typical deletion and include *GTF2I* (or *TFII-I*), *GTF2IRD1* (also called *GTF3*, *MusTRD1*, *BEN*, *CREAM*, *WBSCR11*), and *GTF2IRD2* (Tipney et al. 2004). These genes are the main candidates for the pathology of WBS, in particular, the WBSCP.

Of the TFII-I family, *GTF2I* was the first member identified and is, therefore, best understood (Cheriyath and Roy 2001). GTF2I contains an LZ, which is essential for homodimerization, and six I-repeats. The I-repeats are specific to TFII-I family members, and have a putative helix-loop-helix (HLH) structure (Roy et al. 1997). Dimerization is thought to be facilitated through the I-repeats (Cheriyath and Roy 2001) in a manner similar to known dimeric transcription factors that contain both HLH and LZ structures (e.g., Max of the basic/HLH/Z transcription factors; Ferre-D'Amare et al. 1993). Structural similarities between GTF2I and characterized b/HLH/Z proteins, such as the presence of a DNA binding domain upstream of the second I-repeat, also suggest a comparable mechanism of DNA binding (Roy et al. 1991; Ferre-D'Amare et al. 1993). GTF2I is, however, unusual as a transcription factor in that it appears to act as both a basal factor through binding the transcription start site initiator element (Inr) and as a signal-inducible factor through binding Inr or E-box elements at enhancers (Roy et al. 1997).

Less is known about GTF2IRD1. It contains an LZ and five I-repeats, with an additional sixth I-repeat present in some isoforms of the mouse ortholog. Similar to GTF2I, the LZ is essential for homodimerization (Tantin et al. 2003; Vullhorst and Buonanno 2003). Heterodimers between GTF2I and GTF2IRD1 are not thought to occur even though both proteins contain an LZ and I-repeats (Cheriyath and Roy 2001; Vullhorst and Buonanno 2003). A transcriptional activation domain has been localized to the N terminus of GTF2IRD1 (Yan et al. 2000), and a DNA binding region has been located in the fourth I-repeat of GTF2IRD1 (Vullhorst and Buonanno 2003). GTF2IRD1/Gtf2ird1 is capable of both repressing and activating transcription from a single promoter (Tantin et al. 2003), and appears to be involved in protein interactions (e.g., retinoblastoma protein) (Yan et al. 2000).

GTF2IRD2 is the most recent addition to the family (Tipney et al. 2004), and is located in the repetitive LCR region. Unlike the other members, GTF2IRD2 has a C-terminal CHARLIE8 transposon-like domain, thought to be the consequence of a random in-frame insertion of a transposable element generating a fusion gene, while the N terminus resembles the TFII-I family, containing both a putative LZ and two I-repeats. Its function has yet to be defined.

In this article we compare the TFII-I family members in light of the recently identified GTF2IRD2, and make functional predictions using bioinformatics analysis tools. Similarities and differences are identified that allow putative annotation of DNA binding residues, and based on our motif searches, several potential mechanisms for the regulation of TFII-I proteins are proposed. These include PEST sequences, a protein sequence enriched in proline, glutamate, serine, and threonine known to target proteins for degradation by the 26S proteasome (Rechsteiner and Rogers 1996), synergy control motifs (SCMs) and sumoylation sites (SUMO) that could be involved in regulating the stability and transcriptional activity of these factors. The identification of novel truncated isoforms for each family member is also reported, and suggests similar mechanisms of regulation exist, some of which indicate crossregulation between family members.

Results

Phylogenetic analysis

Figure 1A shows a schematic of the TFII-I family members identified to date. A phylogenetic study of the TFII-I protein family was undertaken to evaluate the genetic variability of the I-repeats both within and between the family members and to estimate phylogenetic relationships between them and their mouse orthologs.
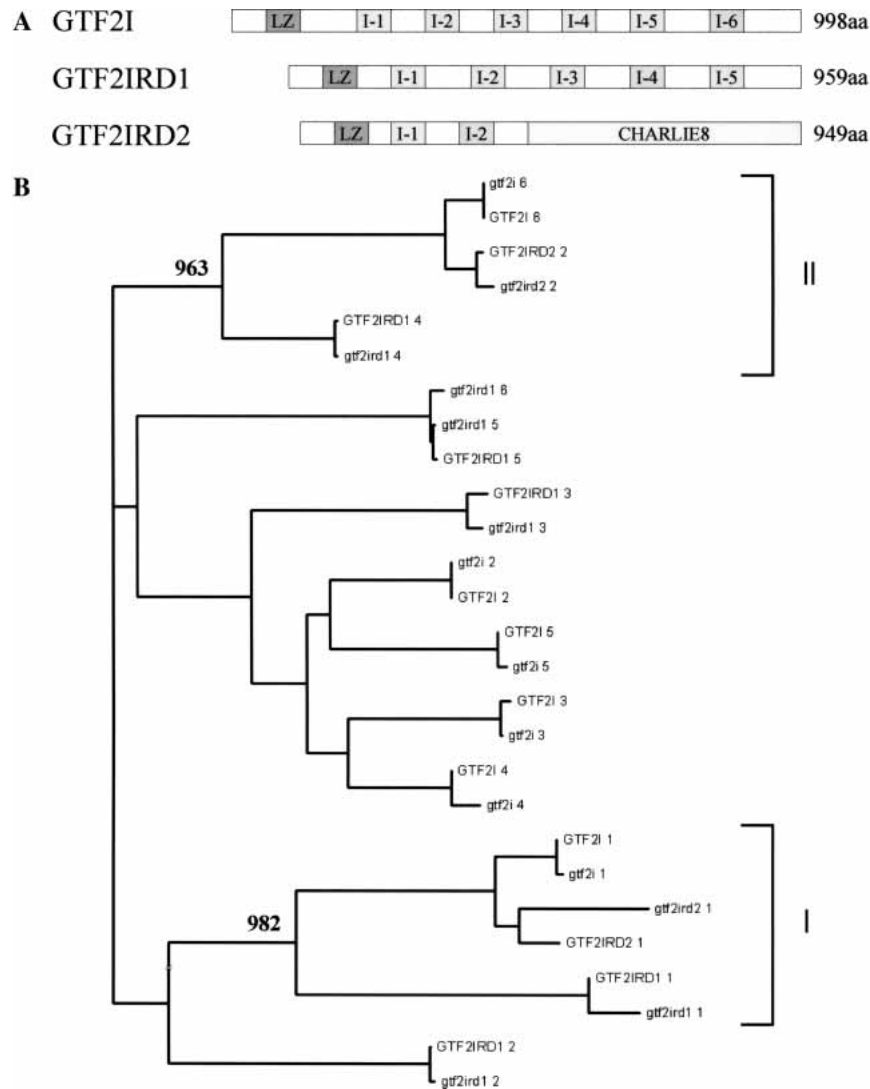
Figure 1. (*A*) Schematical representation of TFII-I family members. LZ (dark gray) and I-repeat structural elements are shown. I-repeats are patterned in accordance to the clade groupings identified through phylogenetic analysis, (see *below*) and the length of each protein is given. (*B*) Phylogeny of I-repeats. Neighbor-joining tree of I-repeats found in the three members of the TFII-I family; GTF2I, GTF2IRD1, and GTF2IRD2. Amino acid sequences for the full-length proteins for both human and mouse (lowercase) I-repeats are detailed. Bootstrap values calculated for 1000 replicates are shown for clades I and II, and the scale bar indicates distance of divergence.

The neighbor-joining method (Saitou and Nei 1987) was used to construct the human/mouse I-repeat tree (Fig. 1B). Each I-repeat clusters with its ortholog except for mouse Gtf2ird1 R-6 (I-repeat 6), which does not have a human ortholog. The first I-repeats (aligned from N to C terminus) of TFII-I family members cluster together with a highly significant bootstrap value (982/1000 replicates) in a clade designated I. GTF2I/Gtf2i R-6, GTF2IRD1/Gtf2ird1 R-4, and GTF2IRD2/Gtf2ird2 R-2 cluster, together with a high bootstrap value (963/1000 replicates). This clade is designated II.

Structure/function prediction

Leucine zipper motif

The LZ regions from human TFII-I family members are aligned in Figure 2. LZs are constructed from heptad repeats with each amino acid position designated **a–g**. TFII-I family LZs contain three heptad repeats. Rules of residue usage in the LZs of B-ZIP proteins that influence hetero/homodimerization are applicable to LZs of all proteins (Vinson et al. 2002). Amino acids at heptad positions **a**, **d**, **e**, and **g** dictate

Heptad position	b	c	d	e	f	g	a	b	c	d	e	f	g	a	b	c	d	e	f	g	a	b	c	d	e	f	g	a
GTF2I LZ	M	V	V	T	F	L	M	S	A	L	E	S	M	C	K	E	L	A	K	S	K	A	E	V	A	C	I	A
GTF2IRD2 LZ	M	V	V	T	F	L	V	S	A	L	E	S	M	C	K	E	L	A	E	S	K	A	E	V	A	C	I	A
GTF2IRD1 LZ	E	I	I	T	S	L	V	S	A	L	D	S	M	C	S	A	L	S	K	L	N	A	E	V	A	C	V	A
			ψ	ψ		ψ	ψ			ψ	-		ψ	ψ			ψ							ψ	ψ		ψ	ψ

Figure 2. Multiple alignment of leucine zipper domains found in the three members of the human TFII-I family. Leucine zipper sequences are aligned with the heptad positions a–g, labeled *above*. Residues at key heptad positions a, d, e, or g (bold) have their conserved physiochemical properties *below* the alignment (ψ, hydrophobic; −, negatively charged residues) and those that are different to their TFII-I family counterparts are highlighted.

LZ oligomerization, dimerization stability, and dimerization specificity; these positions are labeled in Figure 2, and amino acids that differ between TFII-I family members highlighted. Key indicators such as leucine and valine residues at positions **a** and **d**, as well as the negatively charged (glutamic or aspartic acid) residue at the second **e** position are commonly used by LZs to restrict oligomerization to dimers; however, the greater proportion of **a**, **d**, **e**, and **g** positions are hydrophobic, which indicate that these LZ domains could participate in higher order structures (Vinson et al. 2002).

I-repeat motif

Figure 3A shows a multiple alignment of all the human I-repeat sequences from GTF2I, GTF2IRD1, and GTF2IRD2, grouped in accordance with the clades designated in Figure 1B. Residues conserved in all I-repeats as well as hydrophobic positions that are spaced three or four residues apart (characteristic of interacting helices) are incorporated into an I-repeat consensus below the alignment. The HLH consensus of Murre et al. (1989) is aligned against the I-repeat consensus (Fig. 3A). This consensus alignment was manually guided to maximize shared positions but also prioritize key structural residues in the HLH consensus as identified by Ferre-D'Amare et al. (1993). Phenylalanine of the helix 1 consensus is one such residue, bonding to a conserved IL doublet (also conserved in 11 of 13 of the I-repeats) in the helix 2 consensus of the other dimer member. The hydrophobic interactions produced by this bonding are essential to establishing the hydrophobic core of the four-helix

Figure 3. Predicted I-repeat structure. (*A*) Multiple alignment of I-repeats from GTF2I, GTF2IRD1, and GTF2IRD2. Individual I-repeats (grouped in clades) are numbered according to their location from the N terminus of the protein. The HLH consensus (Murre et al. 1989) is manually aligned against the I-repeat consensus: Conserved residues are in bold; partially conserved residues (P,G) in the putative loop region are highlighted gray; ψ denotes conserved hydrophobic positions. The HLH consensus described in Roy et al. (1997) is aligned against the appropriate residues (Φ and Ω are not defined in the original text). The six basic residues required for GTF2I Inr binding are underlined, and putative basic residues required for GTF2IRD1 I-4 DNA binding are boxed. (*B*) Comparison of the putative HLH structure for GTF2I I-R5 against a PSI-PRED secondary structure prediction of the protein (typical of all I-repeats). The residues predicted with helical (cylinder) or β-strand structures (arrow) are highlighted in gray, with random coil sequence represented by a straight line in the cartoon.

bundle (formed on dimerization of HLH proteins; Ferre-D'Amare et al. 1993), and as such, are the key residues that anchor the HLH consensus (Murre et al. 1989) to the I-repeat consensus. Three of the four hydrophobic positions in helix 1 and four of the eight hydrophobic positions in helix 2 of the HLH consensus align to conserved hydrophobic residues found in the I-repeats (Fig. 3A). Such an alignment would predict I-repeats with an HLH structure and a loop of around 40 residues in size. However, several conserved hydrophobic positions in the I-repeats do not have a corresponding position in the HLH consensus. In addition, the original HLH consensus assigned by Roy et al. (1997) is shown in Figure 3A. PSIPRED predicts (Fig. 3B) an extended helix 1 that incorporates additional conserved hydrophobic residues of the I-repeat. Residues toward the C terminus of this helix are assigned a lower confidence. In summary, PSIPRED prediction of I-repeat secondary structure places helices at the positions of the conserved 3/4 spaced hydrophobic residues and predicts that the I-repeats have a long helix 1, with a β-strand and helix in the loop region, in combination with a shorter helix and β-strand predicted at the helix 2 consensus region (see Fig. 3B). Although PSIPRED predicts protein secondary structure with an accuracy of around 75%, it must be noted that predictions of α-helix, β-strand, and random coil structures are indistinguishable from the incorrect 25% (Jones 1999). In addition, the I-repeats are not identified by Pfam Hidden Markov Models generated for the HLH family (data not shown), suggesting I-repeats do not form the structure comparable to this family. If I-repeats do form some HLH structure they are unique in possessing such a large loop (~40 residues), as characterized HLH proteins generally have loops ranging from 5–15 residues and rarely larger than 30 residues (PF00010, pfam helix-loop-helix DNA binding domain; Ferre-D'Amare et al. 1993).

Additional structural features may exist in the I-repeat loop through the inclusion of proline residues. Proline residues can be used as structural pivots in peptides and are often found at turns, usually in combination with glycine residues (MacArthur and Thornton 1991). All proline residues and neighboring glycines within two residues in the predicted loop region of the I-repeat consensus were identified (highlighted gray in Fig. 3A). Analysis showed distinct fingerprints of proline and glycine combinations, which segregate with the clade groupings defined in Figure 1B and might have structural/functional significance.

Functional motif searches

TFII-I family member sequences were searched against multiple resources (PESTfind, SUMOplot, and predictNLS) to detect previously uncharacterized motifs that may have functional significance.

DNA binding sites

Analysis of the I-repeats for key DNA binding residues in GTF2IRD1 I-4 identified a region of basic residues not conserved in all I-repeat members (boxed in Fig. 3A) that is currently undergoing functional analysis. This is conserved amongst the I-repeats that cosegregate into clade II (Fig. 1), and is partially conserved in GTF2IRD1 I-2, also thought to bind DNA (Polly et al. 2003).

Regulatory motifs

PEST (Rechsteiner and Rogers 1996) and SUMO (Kim et al. 2002; Melchior et al. 2003; Seeler and Dejean 2003) sequence predictions were found in all TFII-I family members (Fig. 4A–C). PEST sequences are hydrophilic, and contain proline (P), glutamate (E), serine (S), and threonine (T), usually in a stretch of 12 or more residues, and are commonly flanked by lysine (K), arginine (R), or histidine (H). The minimal consensus target site for SUMO has been identified as ψKxE/D although a few nonconforming conjugation sites are known (TKET and VKYC; Melchior et al. 2003). A synergy control motif (SCM) should contain a sumoylation target site flanked by prolines within four residues up- and downstream from the minimal consensus ($Px_{0-3}\psi KxEx_{0-3}P$). Each of the TFII-I family members contain one SCM (Fig. 4A–C), as well as an inter-I-repeat region containing potential regulatory elements. In GTF2IRD1, this region is between I-2 and I-3, and contains PEST, SCM, and sumoylation sites. In GTF2IRD2, this region is between I-1 and I-2, and similarly contains PEST, SCM, and sumoylation sites. These motifs are localized between I-1 and I-2 in GTF2I, alongside additional important sites that include the nuclear localization signal (NLS), ERK binding domain (residues 323–333 from AC:NP_127492), DNA binding region, and a key phosphorylation target, tyrosine 248 (Y248). It should be noted that in GTF2I, although the key regulatory Y248 is conserved in the mouse, the ERK binding domain is not (Fig. 4A–C).

Shared peptide sequences

Peptide sequences Ccr2, Ccr3, Tcr2, and Tcr3 (Yan et al. 2000) are shared between GTF2I and GTF2IRD1. We detected a similar one to Ccr3/Tcr3 in GTF2IRD2, and designated it Mcr3 based on the amino acid residue difference (Fig. 4C).

Truncated protein isoforms

Through database mining (BLAST and NCBI) and cDNA library screening we have identified short isoforms for all members of the human TFII-I family (Fig. 5). In contrast, mouse database searches only uncovered evidence of a short isoform for the Gtf2ird1 protein. Sequencing of both mouse and human *GTF2IRD1* cDNA clones for the short isoforms
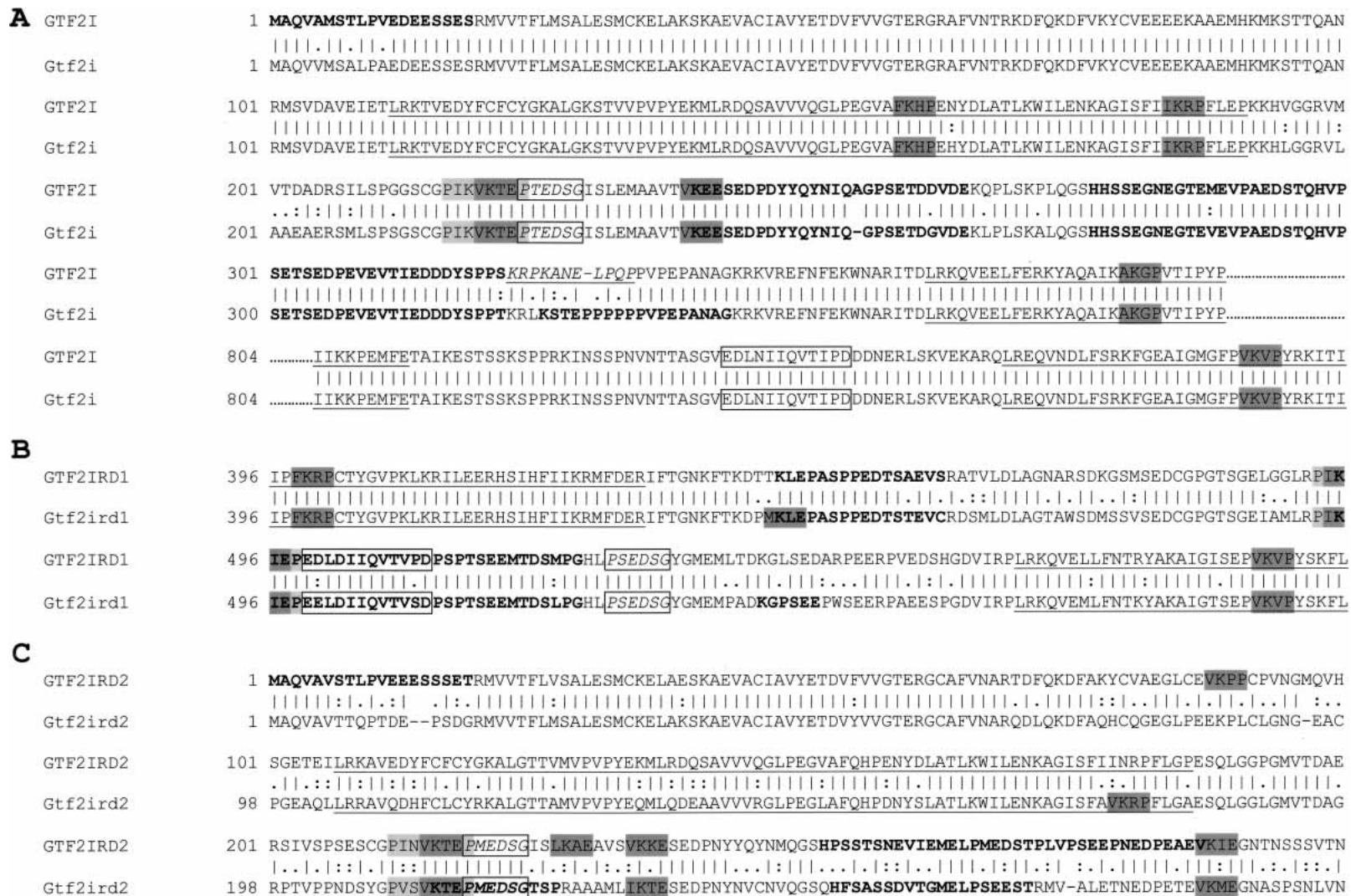
A

```
GTF2I      1 MAQVAMSTLPVEDEESSESRMVVTFLMSALESMCKELAKSKAEVACIAVYETDVFVVGTERGRAFVNTRKDFQKDFVKYCVEEEEKAAEMHKMKSTTQAN
             |||.||.||.|||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Gtf2i      1 MAQVVMSALPAEDEESSESRMVVTFLMSALESMCKELAKSKAEVACIAVYETDVFVVGTERGRAFVNTRKDFQKDFVKYCVEEEEKAAEMHKMKSTTQAN

GTF2I    101 RMSVDAVEIETLRKTVEDYFCFCYGKALGKSTVVPVPYEKMLRDQSAVVVQGLPEGVAFKHPENYDLATLKWILENKAGISFIIKRPFLEPKKHVGGRVM
             |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||:||||||||||||||||||||||||||||:||||:
Gtf2i    101 RMSVDAVEIETLRKTVEDYFCFCYGKALGKSTVVPVPYEKMLRDQSAVVVQGLPEGVAFKHPEHYDLATLKWILENKAGISFIIKRPFLEPKKHLGGRVL

GTF2I    201 VTDADRSILSPGGSCGPIKVKTEPTEDSGISLEMAAVTVKEESEDPDYYQYNIQAGPSETDDVDEKQPLSKPLQGSHHSSEGNEGTEMEVPAEDSTQHVP
             ..:|:|||:|||.||||||||||||||||||||||||||||||||||||||||||| |||||.||||.||||.||||.|||||||||||:||||||||||
Gtf2i    201 AAEAERSMLSPSGSCGPIKVKTEPTEDSGISLEMAAVTVKEESEDPDYYQYNIQ-GPSETDGVDEKLPLSKALQGSHHSSEGNEGTEVEVPAEDSTQHVP

GTF2I    301 SETSEDPEVEVTIEDDDYSPPSKRPKANE-LPQPPVPEPANAGKRKVREFNFEKWNARITDLRKQVEELFERKYAQAIKAKGPVTIPYP..............
             |||||||||||||||||||||||.|:|.:| .|.|||||||||||||||||||||||||||||||||||||||||||||||||||||||
Gtf2i    300 SETSEDPEVEVTIEDDDYSPPTKRLKSTEPPPPPPVPEPANAGKRKVREFNFEKWNARITDLRKQVEELFERKYAQAIKAKGPVTIPYP.

GTF2I    804 ............IIKKPEMFETAIKESTSSKSPPRKINSSPNVNTTASGVEDLNIIQVTIPDDDNERLSKVEKARQLREQVNDLFSRKFGEAIGMGFPVKVPYRKITI
                          |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Gtf2i    804 ............IIKKPEMFETAIKESTSSKSPPRKINSSPNVNTTASGVEDLNIIQVTIPDDDNERLSKVEKARQLREQVNDLFSRKFGEAIGMGFPVKVPYRKITI
```

B

```
GTF2IRD1  396 IPFKRPCTYGVPKLKRILEERHSIHFIIKRMFDERIFTGNKFTKDTTKLEPASPPEDTSAEVSRATVLDLAGNARSDKGSMSEDCGPGTSGELGGLRPIK
              ||||||||||||||||||||||||||||||||||||||||||||..||||||||||||||.||.|.|.::||||.|.||..|:||||||||||||:..||||
Gtf2ird1  396 IPFKRPCTYGVPKLKRILEERHSIHFIIKRMFDERIFTGNKFTKDPMKLEPASPPEDTSTEVCRDSMLDLAGTAWSDMSSVSEDCGPGTSGEIAMLRPIK

GTF2IRD1  496 IEPEDLDIIQVTVPDPSPTSEEMTDSMPGHLPSEDSGYGMEMLTDKGLSEDARPEERPVEDSHGDVIRPLRKQVELLFNTRYAKAIGISEPVKVPYSKFL
              |||:||||||||.|||||||||||||.||||||||||||..|:::.|:|:..|||.|:|.||:||||||||||||:||||:||||||:||||||.||||||
Gtf2ird1  496 IEPEELDIIQVTVSDPSPTSEEMTDSLPGHLPSEDSGYGMEMPADKGPSEEPWSEERPAEESPGDVIRPLRKQVEMLFNTKYAKAIGTSEPVKVPYSKFL
```

C

```
GTF2IRD2    1 MAQVAVSTLPVEEESSSETRMVVTFLVSALESMCKELAESKAEVACIAVYETDVFVVGTERGCAFVNARTDFQKDFAKYCVAEGLCEVKPPCPVNGMQVH
              ||||||:|.|..:|     .|:.|.||||||:|||||||||:||||||||||||||:|||||||||||:||:|||||:|||:||||:||..|||.|.||..|| :..
Gtf2ird2    1 MAQVAVTTQPTDE--PSDGRMVVTFLMSALESMCKELAKSKAEVACIAVYETDVYVVGTERGCAFVNARQDLQKDFAQHCQGEGLPEEKPLCLGNG-EAC

GTF2IRD2  101 SGETEILRKAVEDYFCFCYGKALGTTVMVPVPYEKMLRDQSAVVVQGLPEGVAFQHPENYDLATLKWILENKAGISFIINRPFLGPESQLGGPGMVTDAE
              .||.:::|:||:|||.||.|||||.|||||||||:|||:|::|||:|||||||:|||:||||||||||||||||||:.:||||||||..|:||.||||||||.
Gtf2ird2   98 PGEAQLLRRAVQDHFCLCYRKALGTTAMVPVPYEQMLQDEAAVVVRGLPEGLAFQHPDNYSLATLKWILENKAGISFAVKRPFLGAESQLGGLGMVTDAG

GTF2IRD2  201 RSIVSPSESCGPINVKTEPMEDSGISLKAEAVSVKKESEDPNYYQYNMQGSHPSSTSNEVIEMELPMEDSTPLVPSEEPNEDPEAEVKIEGNTNSSSVTN
              |..|.|::|.|||::|.||||||||||.|.:|.|.::||||.|:||...|||...|:|||....|.|:||.:||.|:||.:|.:.|||||||.||:.::|:.|
Gtf2ird2  198 RPTVPPNDSYGPVSVKTEPMEDSGTSPRAAAMLIKTESEDPNYNVCNVQGSQHFSASSDVTGMELPSEESTRMV-ALETNEDPETEVKMEGNASPSNLVN
```

Figure 4. (*A*) Pairwise alignment of GTF2I (NP_127492) and Gtf2i (AAK49788) residues 1–389 and 804–899. Tcr2 site (boxed), Tcr3 site (boxed and italicized text), and the ERK binding site identified in GTF2I (italicized and underlined). (*B*) Pairwise alignment of GTF2IRD1 (AAF19786) and Gtf2ird1 (AAF78367) residues 396–595. Ccr2 site (boxed), Ccr3 site (boxed and italicized text). (*C*) Pairwise alignment of GTF2IRD2 (NP_775808) and Gtf2ird2 (AAG41674) residues 1–300. Mcr3 site (boxed and italicized text). Common annotations: I-repeats (underlined), predicted PEST sequence (bold), sumoylation sites of high probability (dark gray), and synergy control motifs (dark gray sumoylation site flanked by light gray sequence that includes a proline).
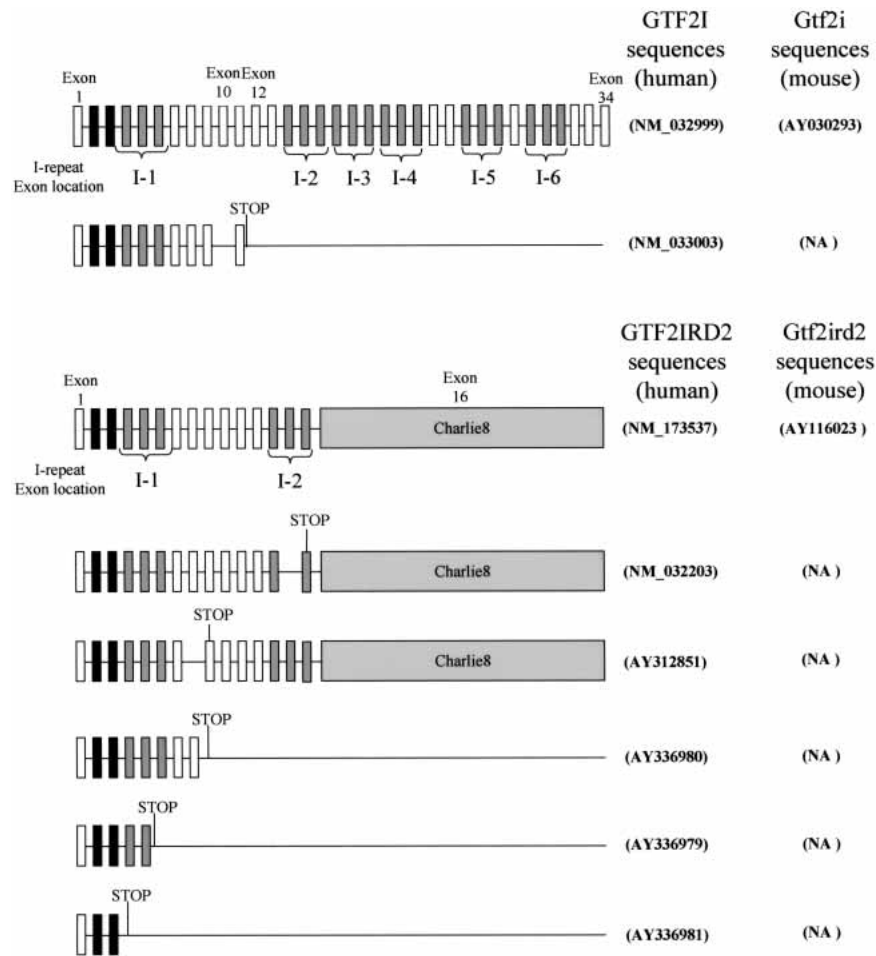
Figure 5. Schematic diagrams of GTF2I/Gtf2i, GTF2IRD2/Gtf2ird2, full-length, and short isoforms. Exons are represented as rectangles. Exons coding for leucine zipper are shaded black; I-repeats are shaded dark gray, with the respective annotated I-repeat number; and GTF2IRD2 charlie8 exon 16 is in light gray. The position of the new termination codon is indicated in the short isoforms. Genbank accession numbers are indicated.

(GenBank BU704306 and AA653462, respectively) showed that they were not full length (no 5′UTR sequence). Although incomplete, the GTF2IRD1 short isoform is predicted to terminate in intron 6 (just after I-2), while the Gtf2ird1 short isoform is predicted to terminate in exon 16 (at the beginning of I-3; data not shown). Figure 5 shows the short isoforms we identified for human GTF2I and GTF2IRD2 and their full-length human and mouse orthologs. Except for Gtf2ird1, the short isoforms terminate before any of the known/predicted DNA binding regions identified to date (Cheriyath and Roy 2001; Vullhorst and Buonanno 2003) and appear to lack a functional nuclear localization signal (NLS). The GTF2I isoform terminates before the known NLS (residues 319–325; Cheriyath and Roy 2000) while PredictNLS analysis of GTF2IRD2 did not identify a NLS throughout the sequence. High-protein sequence homology exists between GTF2I and GTF2IRD2 the N terminus of GTF2IRD2 (residues 1–410) shares 75%

identity with GTF2I and resembles a truncated GTF2I protein (Tipney et al. 2004). The short GTF2IRD2 isoforms are, therefore, almost identical to the short isoform of GTF2I.

Discussion

The results of I-repeat phylogenetic analysis are in agreement with Bayarsaihan et al. (2002), who propose independent duplication histories for I-repeats in GTF2I and GTF2IRD1. Full-length protein alignments suggest that GTF2IRD2 is a truncated version of GTF2I, resembling exons 2–9, 11–12, and 28–31 (Fig. 5) so the GTF2IRD2 I-repeats cluster with GTF2I I-repeats encoded in these exons (Fig.1). Clade I indicates that all TFII-I family members share the N-terminal LZ, and have also retained the equivalent first I-repeat at this position. Whether this is an artifact of the duplication history of the genes/I-repeats or whether

any functional significance can be attributed to this common I-repeat is not known. Clade II is of particular interest because it contains the I-repeat of GTF2IRD1 involved in DNA binding and has a significant bootstrap value. If the functional ability to bind DNA has been retained within this clade then GTF2I I-6 and GTF2IRD2 I-2 could be implicated in DNA binding. Indeed, an E-box binding domain has been described in GTF2I, but has yet to be characterized (Roy et al. 1991).

The assignment of an HLH-like structure to the I-repeats was initially proposed through alignment of this motif in GTF2I with HLH proteins Myc and USF (Roy et al. 1997). The fact that both Myc and USF physically interact with GTF2I and a polyclonal antibody to USF cross-reacts with GTF2I could suggest that similar structures are shared between these proteins (Roy et al. 1991, 1993a). However, the consensus produced from the alignment of Myc, USF, and GTF2I (Roy et al. 1997) does not resemble the HLH family consensus of Murre et al. (1989), to which both Myc and USF belong. Our search for HLH signatures in I-repeats based on the latter HLH consensus alignment also show that, although some hydrophobic positions are comparable, not all the positions have a respective partner so the I-repeats do not conform to the HLH consensus of characterized proteins. The conserved residues of the I-repeat in the proposed first helix are spaced in a 3/4 patterning, which would give it amphipathic properties typical of helices that participate in higher order structures such as the four-helix bundle found in HLH dimers. The second region of conserved residues in the I-repeats fall into two distinct blocks: (1) The first four residues have a 3/4 spacing that could create an amphipathic helix required in the four helix bundle; (2) the last four are compacted over five residue positions and are predicted to form a β-strand. If I-repeats do form a HLH structure, these amphipathic regions may be involved in key interactions. Other interesting predictions are the novel β-strands and a third helix, particularly since the helix and one β-strand are located in the proposed loop region. TFII-I family members are known to function as dimers, and experimental evidence supports that I-repeats are dimerization domains (Cheriyath and Roy 2001). This implies that functional β strands predicted in the I-repeat could form parallel β sheets with the β-strand of the syntenic repeat. If such bonding between β-strands occurs, the loop region would dimerize at its central point bringing together the proposed DNA binding helices within the loop.

Proline residues are unique in that they restrain the backbone conformation of the peptide they are incorporated in (MacArthur and Thornton 1991). The observation that three conserved prolines in the I-repeat consensus were in the putative loop region prompted the search for additional proline as well as glycine residues in this region. The distinct pattern of proline and glycine combinations segregating with the clade groupings implies slight structural differences exist between different I-repeats at these proline regions, which might translate into functional differences. This could explain why specific I-repeats (e.g., GTF2IRD1 I-4) appear to be sufficient for DNA binding even though the peptides studied contain only the I-repeat (Vullhorst and Buonanno 2003; P. Cunliffe, unpubl.). In support, we have highlighted a basic region within GTF2IRD1 I-4 that has the potential to bind DNA and is not conserved in all the I-repeats (Fig. 3A). Its position at the PSI-PRED predicted central helix as well as its conservation in all members of clade II supports the theory that it may contribute to DNA binding. However, the small size of this motif suggests it imparts specificity, with other residues also contributing to the DNA binding.

The leucine zipper (LZ) is a dimerization motif that may be used for the formation of homodimers or heterodimers. However, although the TFII-I family of proteins contain an LZ, the evidence for heterodimerization is ambiguous. There are reports that GTF2I and GTF2IRD1 and their associated LZ domains cannot heterodimerize (Cheriyath and Roy 2001; Vullhorst and Buonanno 2003) alongside others that show that coimmunoprecipitation of GTF2IRD1 (through direct or indirect interactions with GTF2I) is severely hindered if the LZ of GTF2IRD1 is mutated (Tantin et al. 2003). Our analysis of the LZ region has highlighted differences between GTF2I and GTF2IRD1 likely to have a major influence on their interaction potential. For example, the asparagine residue at the third **a** position in GTF2IRD1 is important because this residue is used extensively in B-ZIP proteins to establish homodimerization (Vinson et al. 2002). In comparison, the lysine residues that replace asparagine at this **a** postion in GTF2I and GTF2IRD2 have a destabilizing effect on homodimerization. The only difference at **a**, **d**, **e**, and **g** heptad positions between GTF2I and GTF2IRD*2* LZ is at the first **a** position where methionine replaces valine. Because the hydrophobic property is retained at this site, the similarities between LZs of GTF2I and GTF2IRD2 suggest that heterodimers could form between these two proteins. It is unlikely that GTF2I and GTF2IRD1 heterodimerize, but instead, interact as homodimers. The high proportion of hydrophobic residues in their LZs may contribute to such interactions.

The ubiquitin–proteasome pathway plays a central role in degradation of short-lived and regulatory proteins important in a variety of basic cellular processes. The pathway employs an enzymatic cascade by which multiple ubiquitin molecules are covalently attached to the protein substrate (ubiquitination), thereby marking the protein for destruction and directing it to the 26S proteasome complex for degradation. SUMO (small ubiquitin-like molecule) conjugation is a function specific to eukaryotes, and both the molecule and its mechanism of conjugation are evolutionarily related to ubiquitination. As with ubiquitination, it is the lysine residue in the target site to which SUMO is conjugated. The effect of SUMO conjugation to a protein can vary, and

documented responses range from altering intracellular localization, protein–protein, or protein–DNA interactions, transcription activation abilities and, through synergy control motifs (SCMs), synergism by multiple transcription activators is suppressed at compound promoters (Iniguez-Lluhi and Pearce 2000; Kim et al. 2002; Melchior et al. 2003; Seeler and Dejean 2003; Subramanian et al. 2003). Motif searches on the TFII-I family members highlighted PEST (Rechsteiner and Rogers 1996) and SUMO sequences that may be key in regulation. PEST sequences target proteins for regulated degradation. No universal secondary structure has been observed or predicted for functional PEST sequences but, due to their hydrophilic nature, they are thought to function on the exterior of proteins. PEST-regulated degradation can be initiated by a wide variety of conditional signals, for example, light increases degradation of phytochrome, phosphorylation triggers IκBα degradation, and cAMP binding to cAMP-dependent kinase results in increased degradation (Rechsteiner and Rogers 1996). Because conformational/structural changes are observed in both the phytochrome and cAMP-dependent kinase mechanisms, it is possible that these signals all result in the exposure of previously buried PEST sequences. Evidence from the literature supports that PEST containing target proteins are degraded by the 26S proteasome and, although not all PEST sequences require ubiquitination, most studies suggest prior ubiquitination or a ubiquitin pathway must be functional for PEST-regulated degradation. Functionally key residues within PEST sequences include some TP/SP and K residues. TP/SP residue combinations are potential phosphorylation sites, and may be the key target site that initiates degradation, whereas lysine residues are important because they are the target molecules for ubiquitin conjugation.

SCMs were found in TFII-I family members which, if functional, would implicate them not only as traditional transcription factors, but also with an inbuilt repressive mechanism to limit transcriptional synergism at compound promoters. The minimal sumoylation consensus site is only four residues long, and will arise by chance frequently in a protein sequence. It is, therefore, reasonable to assume that the majority of predicted sites are not functional. However, sites at the edge of PEST sequence are interesting because they could be involved in regulating TFII-I protein family degradation. Because both PEST/ubiquitin-regulated degradation and sumoylation occur at lysine residues, antagonistic competition between ubiquitination and sumoylation at lysine residues that are part of both PEST and SUMO sites could well determine the fate of a protein containing these sites. Support for this comes from the proteins IκBα (an inhibitor molecule of NF-κB) and Mdm2 (an E3 ubiquitin ligase for the p53 tumor suppressor protein). Although no PEST sequence is predicted to exist at their sumoylation sites, these proteins possess lysine residues involved in

ubiquitin/SUMO competition, which dictates degradation fate (Kim et al. 2002). The PEST/SUMO sites detected in the TFII-I family could indicate similar regulatory mechanisms. GTF2I/Gtf2i possess one SUMO site that overlaps with PEST sequence, suggesting this site would dictate degradation while the SCM solely regulates synergy. PEST sequence in Gtf2ird1 overlaps both the SCM and a minimal SUMO site, suggesting both sites could regulate degradation, with the SCM also controlling synergy. GTF2IRD1 only has its SCM overlapping PEST sequence, suggesting this site regulates both synergy and degradation. GTF2IRD2/Gtf2ird2 is interesting, because the mouse sequence has an SCM and a separate PEST/SUMO site, suggesting that two different sites regulate synergy and degradation, respectively. In the human sequence, no PEST sequence is predicted at the minimal SUMO site, but instead, PEST sequence now overlaps the SCM, suggesting this one site regulates both synergy and degradation. This may point to a compensatory mutation between human and mouse PEST sequences, further supporting the hypothesis these PEST/SUMO border lysine residues are functional. Evidence from the literature does suggest that GTF2I and GTF2IRD1 are sumoylation targets, because both the enzyme responsible for sumoylating proteins (Ubc9) and a known catalyst of sumoylation (PIASxβ) have been proven to interact with these TFII-I family members (Tussie-Luna et al. 2002).

A key site conserved in orthologs of GTF2I is the tyrosine residue at position 248 (Y248), whose function is critical because its mutation ablates transcriptional activation (Novina et al. 1999). Phosphorylation of Y248 is known to result in the exposure of a previously buried ERK binding domain, thereby promoting the interaction of GTF2I with ERK (Kim and Cochran 2001) with a consequent increased likelihood of phosphorylating key ERK target sites in GTF2I required for transactivation. However, because this ERK binding domain is currently only found in the human ortholog, the unmasking of this domain could well be human specific, and a secondary artefact of the true role of Y248 phosphorylation, which could be the induction of a conformational change that exposes the regulatory sequence hotspot found between I-1 and I-2. Not only is the ERK binding domain located here, but also the conserved PEST/SCM/sumoylation sites, NLS, DNA binding residues, and Ccr3. Therefore, Y248 phosphorylation could act as a control point not just for transactivation but also synergy, protein stability and protein–protein interactions.

We have also identified very short truncated isoforms of TFII-I family members, which have the potential to play a role as natural antagonists to the full-length protein through dimerization, thereby regulating function. Mechanisms exist in the Id and extramacrochaetae proteins of the HLH transcription factor family where such proteins, resembling truncated isoforms lacking a basic DNA binding region,

form heterodimers to prevent DNA binding of basic HLH proteins (Benezra et al. 1990; Ellis et al. 1990). Although all of the short isoforms (except Gtf2ird1) lack any of the known/predicted DNA binding regions, there is substantial evidence to suggest that the TFII-I members can bind DNA as monomers in vitro (Roy et al. 1991; Tantin et al. 2003). This suggests that these short isoforms might not affect DNA binding directly. An alternative mechanism of action might be to restrict the formation of specific higher order complexes that only arise through interactions with the full-length dimers. The lack of an NLS in these short isoforms suggests that they reside in the cytoplasm only gaining nuclear entry when dimerized with a full-length protein with a functional NLS. The high homology shared between GTF2I and the N terminus of GTF2IRD2 (Tipney et al. 2004), implies that their respective short isoforms would be almost identical and, as such, could well have overlapping functions with both homodimers and heterodimers forming between them and the full-length GTF2I and GTF2IRD2 proteins. This predicted cross-reactivity might have been the evolutionary driving force behind the chromosomal re-arrangement at 7q11.23 that is not found in the syntenic region in mouse. The rearrangement has resulted in the duplication of *GTF2IRD2* on chromosome 7, with both copies producing short isoforms (Tipney et al. 2004). This may bestow advantageous properties in humans such as greater protein levels or differential protein expression through the use of different promoter/regulatory elements, allowing greater regulation through these short isoforms.

Conclusion

The emerging picture describing the TFII-I transcription factor proteins is one of a family of scaffolding/architectural transcription factors. Several lines of evidence support such a claim. First, *GTF2I* was identified as a basal transcription factor, and has been shown to interact with a number of proteins such as Phox1, SRF, USF1, c-Myc, STAT1, STAT3, ATF6, NFkβ, HDAC3, and PIASxβ (Roy et al. 1991, 1993a,b; Montano et al. 1996; Grueneberg et al. 1997; Kim et al. 1998; Parker et al. 2001; Roy 2001). Although *GTF2IRD1* was the second member of the family identified, knowledge of its interacting partners is increasing (e.g., HDAC3, PIASxβ, and Rb). Second, GTF2IRD1 has been shown to have both repressive and activation properties at the same promoter in different cell types (Tantin et al. 2003). This suggests these features may not be an intrinsic property of these proteins but are the result of context (promoter/cell-type/signal)-dependent influences that dictate the state and which cofactors interact with these proteins at a given location and time point. Third, a GTF2I construct (p70) lacking the C-terminal 222 amino acids has been shown to bind the Inr element, but is unable to activate reporter genes via this element (Cheriyath et al. 1998). It

was therefore thought that like traditional transcription factors the N terminus contained the DNA binding domain, while the C terminus contained the activation domain. However, a fusion construct made between a GAL4 DNA binding domain and the proposed GTF2I C-terminal activation domain was not functional, suggesting either the activation domain of GTF2I is not modular, or requires posttranslational modification, or that the activation properties of GTF2I are gained through recruiting another protein that cannot bind simply to sequence found at the C terminus. Finally, in the studies of Polly et al. (2003), GTF2IRD1 binding at the USE Troponin I slow enhancer was abolished through promoter mutation. This mutation almost completely eliminated any basal transcription, demonstrating that, even though the GTF2IRD1 binding site is classified as an enhancer, it mediates basal activity and, as such, is an essential component of the USE Troponin I promoter.

The detailed assessment of the I-repeat secondary structure identified conserved proline/glycine fingerprints that will have distinct structures and two predicted amphipathic helices characteristic of those involved in higher order structures. Although these I-repeats may form an HLH structure, they do not conform to the consensus of characterized HLH proteins, and would possess an unusual large loop of around 40 residues. Although the I-repeats were initially proposed to interact with proteins, no evidence, discounting interactions between I-repeats during homodimerization, exists to date. Preliminary data are, however, available to show that they are involved in DNA binding (Polly et al. 2003; Vullhorst and Buonanno 2003; P. Cunliffe, unpubl.). GTF2I I-4 has been highlighted as interacting with cGMP-dependent Protein Kinase Iβ, although on closer examination the residues downstream of I-4 actually perform the majority of this interaction (Casteel et al. 2002). Protein–protein interactions that might be universal to all members of the TFII-I family have been described (e.g., HDAC3 and PIASxβ bind both GTF2I and GTF2IRD1; Tussie-Luna et al. 2002), which may infer common binding sites retained in this family of proteins such as the similar peptide sequences Ccr2, Ccr3, Tcr2, Tcr3 (Casteel et al. 2002), and Mcr3.

We have detailed several potential mechanisms indicating how the TFII-I proteins may be regulated individually and as a family. PEST, SCM, and SUMO sites identified in the protein sequences point to a common regulatory pathway. Evidence for the existence of short isoforms for all human TFII-I members also indicate shared regulatory mechanisms, with high sequence similarity between GTF2I and GTF2IRD2 short isoforms predicting cross-reaction between them. GTF2I and GTF2IRD1 have recently been shown to interact (Tantin et al. 2003) and, when coexpressed, GTF2IRD1 is known to influence GTF2I nuclear localization (Tussie-Luna et al. 2001). The Immunoglobulin promoter is bound and regulated by both GTF2I and

GTF2IRD1 (Tantin et al. 2003). All of these proven and predicted interactions point to a complex network that regulates the TFII-I family of proteins and their target genes.

The TFII-I family of transcription factors clustering at 7q11.23, therefore, appear to be complex multifunctional proteins, and their emerging properties alongside patient data implicates them in the main pathology of WBS, especially the WBSCP neurodevelopmental phenotypes. However, given the complicated interactions that could occur between the family members, the unambiguous identification of individual protein contributions to these phenotypes as a result of haploinsufficiency presents an intricate task.

Materials and methods

The following sequences were used in analysis: GTF2I (NP_127492), *Gtf2i* (AAK49788), *GTF2IRD1* (AAF19786), *Gtf2ird1* (AAF78367), *GTF2IRD2* (NP_775808), and *Gtf2ird2* (AAG41674).

Bioinformatics tools

Phylogenetic trees

The 90 amino acids designated to comprise the I-repeat domain (Roy et al. 1997) were collated from all human and mouse sequences of the TFII-I family and aligned using CLUSTALX (version 1.8; Thompson et al. 1997). The output PHY file produced from this program was used in the PHYLip (version 3.6a; Felsenstein 1989) collection of programs to create a phylogenetic tree. PROTDIST was used to calculate evolutionary distance matrices using the Dayhoff PAM matrix settings. The neighbor-joining method was used to construct trees. Bootstrap values were calculated from 1000 replicates using SeqBoot.

Secondary structure prediction

Prediction of the secondary structure was performed using the PSIPRED protein structure prediction server (McGuffin et al. 2000; http://bioinf.cs.ucl.ac.uk/psipred/).

Motif searches

PEST sequences were searched using PESTfind (http://www.at.embnet.org/embnet/tools/bio/PESTfind), while sumoylation sites were searched for using SUMOplot (http://www.abgent.com/sumoplot.html).

Short isoform sequence searches

Each of the TFII-I family sequences (accession numbers above) were used in BLAST searches (Altschul et al. 1990) against the Genbank database (http://www.ncbi.nlm.nih.gov) to identify truncated isoforms of the family members.

Additional tools

CLUSTALW (http://www.ebi.ac.uk/clustalw) is a multiple alignment tool that was utilized to produce such alignments of I-repeats from all human TFII-I members.

Isolation of GTF2IRD2 truncated isoforms

GTF2IRD2 clones were isolated from an 18-week human fetal brain cDNA library (Gibco) screened with probes made from the mouse *Gtf2ird2* cDNA. Hybridization conditions were according to the manufacturer's instructions. Positive cDNA clones were sequenced by fluorescent BigDye terminator cycle sequencing (V 2.0 kit, Applied Biosystems) using T7 and T3 vector primers, and visualized on an ABI 373 sequencer. cDNA clones were also obtained from the IMAGE consortium (HGMP UK) and sequenced. Sequences were submitted to Genbank (AY312851, splice variant 1; AY336979, GTF2IRD2 isoform 2; AY336981, GTF2IRD2 isoform 3; AY336980, GTF2IRD2 isoform 1).

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Baumer, A., Dutly, F., Balmer, D., Riegel, M., Tukel, T., Krajewska-Walasek, M., and Schinzel, A. 1998. High level of unequal meiotic crossovers at the origin of the 22q11.2 and 7q11.23 deletions. *Hum. Mol. Genet.* **7:** 887–894.

Bayarsaihan, D., Dunai, J., Greally, J.M., Kawasaki, K., Sumiyama, K., Enkhmandakh, B., Shimizu, N., and Ruddle, F.H. 2002. Genomic organization of the genes Gtf2ird1, Gtf2i and Ncf1 at the mouse chromosome 5 region syntenic to the human chromosome 7q11.23 Williams syndrome critical region. *Genomics* **79:** 137–143.

Bellugi, U., Lichtenberger, L., Jones, W., and Lai, Z. 2000. The neurocognitive profile of Williams syndrome: A complex pattern of strengths and weaknesses. *J. Cogn. Neurosci.* **12 (Suppl. 1):** 7–29.

Benezra, R., Davis, R.L., Lockshun, D., Turner, D.L., and Weintraub, H. 1990. The protein Id: A negative regulator of helix-loop-helix DNA binding proteins. *Cell* **61:** 49–59.

Bennett, C., La Veck, B., and Sells, C.J. 1978. The Williams elfin facies syndrome: The psychological profile as an aid in syndrome identification. *Pediatrics* **61:** 303–306.

Beuren, A.J., Apitz, J., and Harmjanz, D. 1962. Supravalvular aortic stenosis in association with mental retardation and a certain facial appearance. *Circulation* **26:** 1235–1240.

Botta, A., Novelli, G., Mari, A., Novelli, A., Sabani, M., Korenberg, J.R., Osborne, L.R., Digilio, M.C., Giannotti, A., and Dallapiccola, B. 1999. Detection of an atypical 7q11.23 deletion in Williams syndrome patients which does not include STX1A and FZD3 genes. *J. Med. Genet.* **36:** 478–480.

Casteel, D.E., Zhuang, S., Gudi, T., Tang, J., Vuica, M., Desiderio, S., and Pilz, R.B. 2002. cGMP-dependent protein kinase Ib physically and functionally interacts with the transcriptional regulator TFII-I. *J. Biol. Chem.* **277:** 32003–32014.

Cheriyath, V. and Roy, A.L. 2000. Alternatively spliced isoforms of TFII-I. *J. Biol. Chem.* **275:** 26300–26308.

———. 2001. Structure–function analysis of TFII-I. *J. Biol. Chem.* **276:** 8377–8383.

Cheriyath, V., Novina, C.D., and Roy, A.L. 1998. TFII-I regulates Vβ promoter activity through an initiator element. *Mol. Cell. Biol.* **18:** 4444–4454.

Cheriyath, V., Desgranges, Z.P., and Roy, A.L. 2002. c-Src-dependent transcriptional activation of TFII-I. *J. Biol. Chem.* **277:** 22798–22805.

Curran, M.E., Atkinson, D.L., Ewart, A.K., Morris, C.A., Leppert, M.F., and Keating, M.T. 1993. The Elastin gene is disrupted by a translocation associated with Supravalvular Aortic Stenosis. *Cell* **73:** 159–168.

Donnai, D. and Karmiloff-Smith, A. 2000. Williams syndrome: From genotype through to the cognitive phenotype. *Am. J. Med. Genet.* **97:** 164–171.

Dutly, F. and Schinzel, A. 1996. Unequal interchromosomal rearrangements may result in elastin gene deletions causing the Williams-Beuren syndrome. *Hum. Mol. Genet.* **5:** 1893–1898.

Ellis, H.M., Spann, D.R., and Posakony, J.W. 1990. Extramacrochaetae, a negative regulator of sensory organ development in *Drosophila*, defines a new class of helix-loop-helix proteins. *Cell* **61:** 27–38.

Ewart, A.K., Morris, C.A., Ensing, G.J., Loker, J., Moore, C., Leppert, M., and Keating, M.T. 1993a. A human vascular disorder, Supravalvular Aortic Stenosis, maps to Chromosome 7. *Proc. Natl. Acad. Sci.* **90:** 3226–3230.

Ewart, A.K., Morris, C.A., Atkinson, D.L., Jin, W., Sternes, K., Spallone, P., Stock, A.D., Leppert, M., and Keating, M.T. 1993b. Hemizygosity at the Elastin locus in a developmental disorder, Williams syndrome. *Nat. Genet.* **5:** 11–16.

Felsenstein, J. 1989. PHYLIP-phylogeny inference package (version 3.2). *Cladistics* **5:** 164–166.

Ferre-D'Amare, A.R., Prendergast, G.C., Ziff, E.B., and Burley, S.K. 1993. Recognition of Max of its cognate DNA through a dimeric b/HLH/Z domain. *Nature* **363:** 38–45.

Francke, U. 1999. Williams-Beuren syndrome: Genes and mechanisms. *Hum. Mol. Genet.* **8:** 1947–1954.

Gagliardi, C., Bonaglia, M.C., Selicorni, A., Borgatti, R., and Giorda, R. 2003. Unusual cognitive and behavioural profile in a Williams syndrome patient with atypical 7q11.23 deletion. *J. Med. Genet.* **40:** 526–530.

Grueneberg, D.A., Henry, R.W., Brauer, A., Novina, C.D., Cheriyath, V., Roy, A.L., and Gilman, M. 1997. A multifactorial DNA-binding protein that promotes the formation of serum response factor/homeodomain complexes: Identity to TFII-I. *Genes & Dev.* **11:** 2482–2493.

Hirota, H., Matsuoka, R., Chen, X.N., Salandanan, L.S., Lincoln, A.L., Rose, F.E., Sunahara, M., Osawa, M., Bellugi, U., and Korenberg, J.R. 2003. Williams syndrome deficits in visual spatial processing linked to GTF2IRD1 and GTF2I on Chromosome 7q11.23. *Genet. Med.* **5:** 311–321.

Iniguez-Lluhi, J.A. and Pearce, D. 2000. A common motif within the negative regulatory regions of multiple factors inhibits their transcriptional synergy. *Mol. Cell. Biol.* **20:** 6040–6050.

Jones, D.T. 1999. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292:** 195–202.

Jones, W., Bellugi, U., Lai, Z., and Chiles, M. 2000. Hypersociability in Williams syndrome. *J. Cogn. Neurosci.* **12 (Suppl 1):** 30–46.

Kim, D.W. and Cochran, B.H. 2001. JAK2 activates TFII-I and regulates its interaction with Extracellular Signal-Regulated Kinase. *Mol. Cell. Biol.* **21:** 3387–3397.

Kim, D.W., Cheriyath, V., Roy, A.L., and Cochran, B.H. 1998. TFII-I enhances activation of the c-fos promoter through interactions with upstream elements. *Mol. Cell. Biol.* **18:** 3310–3320.

Kim, K.I., Baek, S.H., and Chung, C.H. 2002. Versatile protein tag, SUMO: Its enzymology and biological function. *J. Cell. Physiol.* **191:** 257–268.

Korenberg, J.R., Chen, X.N., and Hirota, H. 2000. Genome structure and cognitive map of Williams syndrome. *J. Cogn. Neurosci.* **12:** 89–107.

Lupski, J.R. 1998. Genomic disorders: Structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet.* **14:** 417–422.

MacArthur, M.W. and Thornton, J.M. 1991. Influence of proline residues on protein conformation. *J. Mol. Biol.* **218:** 397–412.

McGuffin, L.J., Bryson, K., and Jones, D.T. 2000. The PSIPRED protein structure prediction server. *Bioinformatics* **16:** 404–405.

Melchior, F., Schergaut, M., and Pichler, A. 2003. SUMO: Ligases, isopeptidases and nuclear pores. *Trends Biochem. Sci.* **28:** 612–618.

Montano, M.A., Kripke, K., Norina, C.D., Achacoso, P., Herzenberg, L.A., Roy, A.L., and Nolan, G.P. 1996. NF-kB homodimer binding within the HIV-1 initiator region and interactions with TFII-I. *Proc. Natl. Acad. Sci.* **93:** 12376–12381.

Morris, C.A., Demsey, S.A., Leonard, C.O., Dilts, C., and Blackburn, B.L. 1988. Natural history of Williams syndrome: Physical characteristics. *J. Pediatr.* **113:** 318–326.

Murre, C., McCaw, S.P., and Baltimore, D. 1989. A new DNA binding and dimerization motif in immunoglobulin enhancer binding, Daughterless, MyoD, and myc proteins. *Cell* **56:** 777–783.

Novina, C.D., Kumar, S., Bajpai, U., Cheriyath, V., Zhang, Y., Pillai, S., Wortis, H.H., and Roy, A.L. 1999. Regulation of nuclear localisation and transcriptional activity of TFII-I by Bruton's Tyrosine Kinase. *Mol. Cell. Biol.* **19:** 5014–5024.

Osborne, L.R. 1999. Williams-Beuren syndrome: Unraveling the mysteries of a microdeletion disorder. *Mol. Genet. Metab.* **67:** 1–10.

Osborne, L.R., Herbrick, J.A., Greavette, T., Heng, H.H.Q., Tsui, L., and Scherer, S.W. 1997. PMS2-related genes flank the rearrangement breakpoints associated with Williams syndrome and other diseases on human Chromosome 7. *Genomics* **45:** 402–406.

Osborne, L.R., Li, M., Pober, B., Chitayat, D., Bodurtha, J., Mandel, A., Costa, T., Grebe, T., Cox, S., Tsui, L., et al. 2001. A 1.5 million-base pair inversion polymorphism in families with Williams-Beuren syndrome. *Nat. Genet.* **29:** 321–325.

Parker, R., Phan, T., Baumeister, P., Roy, B., Cheriyath, V., Roy, A.L., and Lee, A.S. 2001. Identification of TFII-I as the endoplasmic reticulum stress response element binding factor ERSF: Its autoregulation by stress and interaction with ATF6. *Mol. Cell. Biol.* **21:** 3220–3233.

Perez Jurado, L.A., Peoples, R.J., Kaplan, P., Hamel, B.C.J., and Francke, U. 1996. Molecular definition of chromosome 7 deletion in Williams syndrome and parent-of-origin effects on growth. *Am. J. Hum. Genet.* **59:** 781–792.

Perez Jurado, L.A., Wang, Y., Peoples, R.J., Coloma, A., Cruces, J., and Francke, U. 1998. A duplicated gene in the breakpoint regions of the 7q11.23 Williams-Beuren syndrome deletion encodes the initiator binding protein TFII-I and BAP-135, a phosphorylation target of BTK. *Hum. Mol. Genet.* **7:** 325–334.

Polly, P., Haddadi, L.M., Issa, L.L., Subramaniam, N., Palmer, S.J., Tay, E.S.E., and Hardeman, E.C. 2003. hMusTRD1a1 represses MEF2 activation of the Troponin I slow enhancer. *J. Biol. Chem.* **278:** 36603–36610.

Rechsteiner, M. and Rogers, S.W. 1996. PEST sequences and regulation by proteolysis. *Trends Biol. Sci.* **21:** 267–271.

Robinson, W.P., Waslynka, J., Bayarsaihan, F., Wang, M., Clark, S., Kotzot, D., and Schinzel, A. 1996. Delineation of 7q11.2 deletions associated with Williams-Beuren syndrome and mapping of a repetitive sequence to within and to either side of the common deletion. *Genomics* **34:** 17–23.

Roy, A.L. 2001. Biochemistry and biology of the inducible multifunctional transcription factor TFII-I. *Gene* **274:** 1–13.

Roy, A.L., Meisterernst, M., Pognonec, P., and Roeder, R.G. 1991. Cooperative interaction of an initiator-binding transcription factor and the helix-loop-helix activator USF. *Nature* **354:** 245–248.

Roy, A.L., Carruthers, C., Gutjahr, T., and Roeder, R.G. 1993a. Direct role for Myc in transcription initiation mediated by interactions with TFII-I. *Nature* **365:** 359–361.

Roy, A.L., Malik, S., Meisterernst, M., and Roeder, R.G. 1993b. An alternative pathway for transcription initiation involving TFII-I. *Nature* **365:** 355–358.

Roy, A.L., Du, H., Gregor, P.D., Novina, C.D., Martinez, E., and Roeder, R.G. 1997. Cloning of an Inr and E-box binding protein, TFII-I that interacts physically and functionally with USF1. *EMBO J.* **16:** 7091–7104.

Saitou, N. and Nei, M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4:** 406–425.

Seeler, J.S. and Dejean, A. 2003. Nuclear and unclear functions of SUMO. *Nature* **4:** 690–699.

Stankiewicz, P. and Lupski, J.R. 2002. Molecular-evolutionary mechanisms for genomic disorders. *Curr. Opin. Genet. Dev.* **12:** 312–319.

Subramanian, L., Benson, M.D., and Iniguez-Lluhi, J.A. 2003. A synergy control motif within the attenuator domain of C/EBPα inhibits transcriptional synergy through its PIASy-enhanced modification by SUMO-1 or SUMO-3. *J. Biol. Chem.* **278:** 9134–9141.

Tantin, D., Tussie-Luna, M.I., Roy, A.L., and Sharp, P.A. 2003. Regulation of immunoglobulin promoter activity by TFII-I class transcription factors. *J. Biol. Chem.* **279:** 5460–5469.

Tassabehji, M., Metcalfe, K., Karmiloff-Smith, A., Carette, M.J., Grant, J., Dennis, N., Reardon, W., Splitt, M., Read, A.P., and Donnai, D. 1999. Williams syndrome: Use of chromosomal microdeletions as a tool to dissect cognitive and physical phenotypes. *Am. J. Hum. Genet.* **64:** 118–125.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., and Higgins, D.G. 1997. The CLUSTAL_X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25:** 4876–4882.

Tipney, H.J., Hinsley, T.A., Brass, A., Metcalfe, K., Donnai, D., and Tassabehji, M. 2004. Isolation and characterisation of GTF2IRD2, a novel fusion gene mapping to the Williams-Beuren syndrome critical region. *Eur J. Hum. Genet.* **12:** 551–560.

Tussie-Luna, M.I., Bayarsaihan, D., Ruddle, F.H., and Roy, A.L. 2001. Repression of TFII-I-dependent transcription by nuclear exclusion. *Proc. Natl. Acad. Sci.* **98:** 7789–7794.

Tussie-Luna, M.I., Bayarsaihan, D., Seto, E., Ruddle, F.H., and Roy, A.L. 2002. Physical and functional interactions of histone deacetylase 3 with TFII-I family proteins and PIASxβ. *Proc. Natl. Acad. Sci.* **99:** 12807–12812.

Udwin, O. and Yule, W. 1991. A cognitive and behavioural phenotype in Williams syndrome. *J. Clin. Exp. Neuropsychol.* **13:** 232–244.

Vinson, C., Myakishev, M., Acharya, A., Mir, A.A., Moll, J.R., and Bonovich, M. 2002. Classification of human B-ZIP proteins based on dimerization properties. *Mol. Cell. Biol.* **22:** 6321–6335.

Vullhorst, D. and Buonanno, A. 2003. Characterisation of general transcription factor 3, a transcription factor involved in slow muscle-specific gene expression. *J. Biol. Chem.* **278:** 8370–8379.

Williams, J.C.P., Barratt-Boyes, B.G., and Lowe, J.B. 1961. Supravalvular aortic stenosis. *Circulation* **24:** 1311–1318.

Yan, X., Zhao, X., Qian, M., Guo, N., and Zhu, X. 2000. Characterization and gene structure of a novel retinoblastoma-protein-associated protein similar to the transcription regulator TFII-I. *Biochem. J.* **345:** 749–757.