Supplementary Data

Structural basis for preferential binding of human TCF4 to DNA containing 5-carboxylcytosine

Jie Yang[1], John R. Horton[1], Jia Li[2], Yun Huang[2], Xing Zhang[1], Robert M. Blumenthal[3], Xiaodong Cheng[1,*]

[1] Department of Molecular and Cellular Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA

[2] Center for Epigenetics & Disease Prevention, Institute of Biosciences and Technology, College of Medicine, Texas A&M University, Houston, TX 77030, USA

[3] Department of Medical Microbiology and Immunology, and Program in Bioinformatics, The University of Toledo College of Medicine and Life Sciences, Toledo, OH 43614, USA

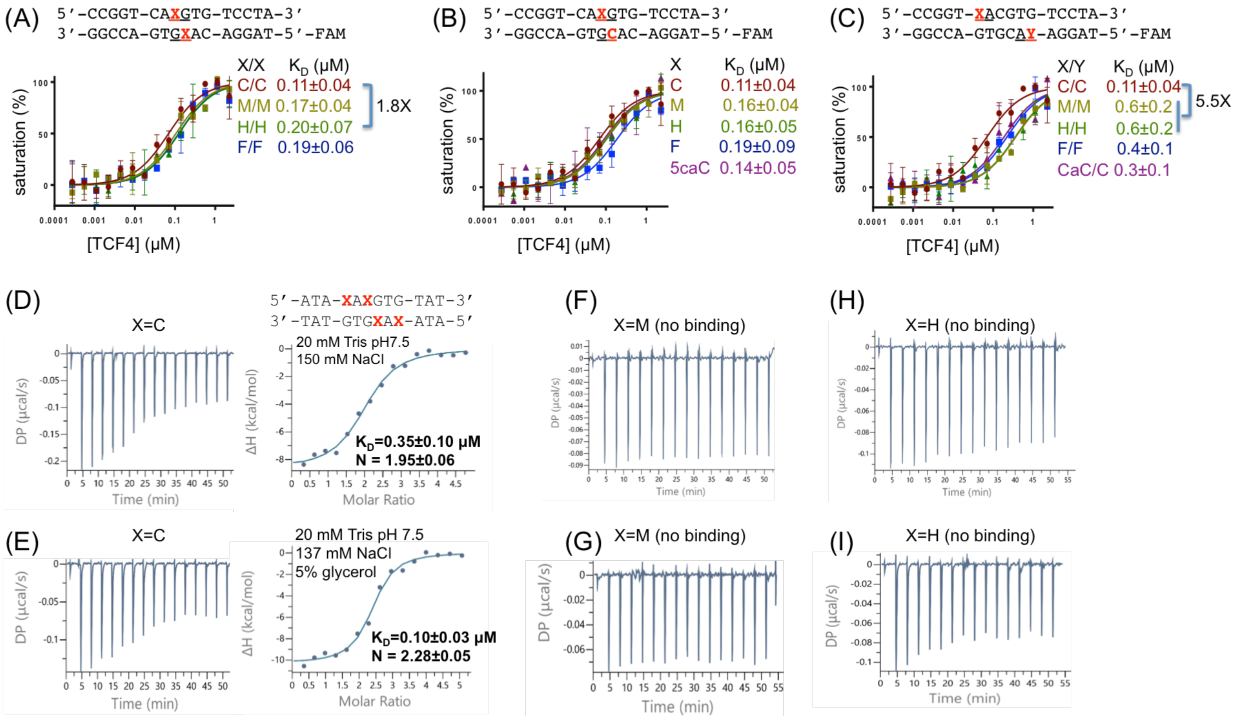
*To whom correspondence should be addressed: Email: xcheng5@mdanderson.org

All authors email addresses:

JY (jyang19@mdanderson.org); JRH (jrhorton@mdanderson.org); JL (jli@ibt.tamhsc.edu); YH (yun.huang@ibt.tamhsc.edu); XZ (xzhang21@mdanderson.org); RMB (robert.blumenthal@utoledo.edu)

6 Supplementary Figures and 1 Supplementary Table

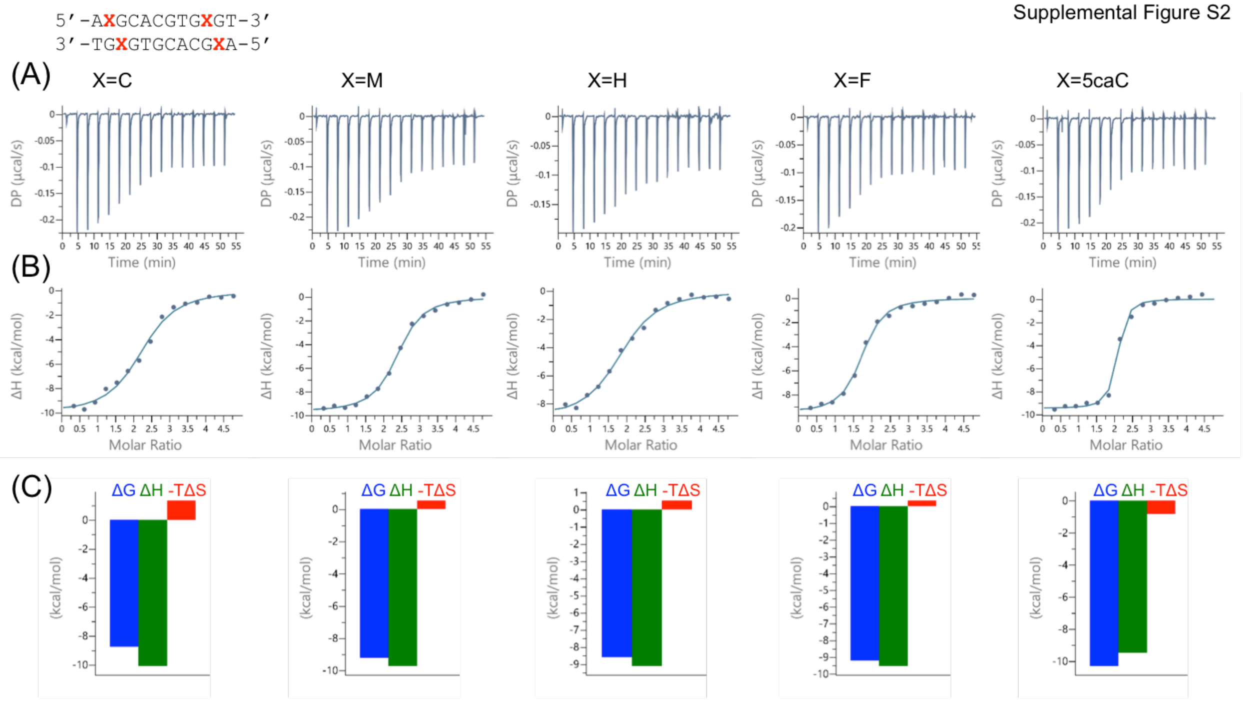
Supplemental Figure S1



Supplementary Figure S1. Related to Table 1A-1D

(A-C) The FP measurements of KD values by various cytosine modifications of oligonucleotides at central CpG site (A and B) or CpA sites (C) against TCF4 bHLH domain. (**D-E**) The ITC measurement of KD values by unmodified oligonucleotide was carried out under the setup conditions of [dsDNA] = 3.3 μM being kept in the sample cell and [TCF4] (monomer) = 77.7 μM being injected into the cell by a syringe under two buffer conditions. The derived K_D values are sensitive to ionic strength and glycerol. (**F-I**) No bindings were observed for modifications (M or H) at all cytosines.

Supplemental Figure S2



(D)

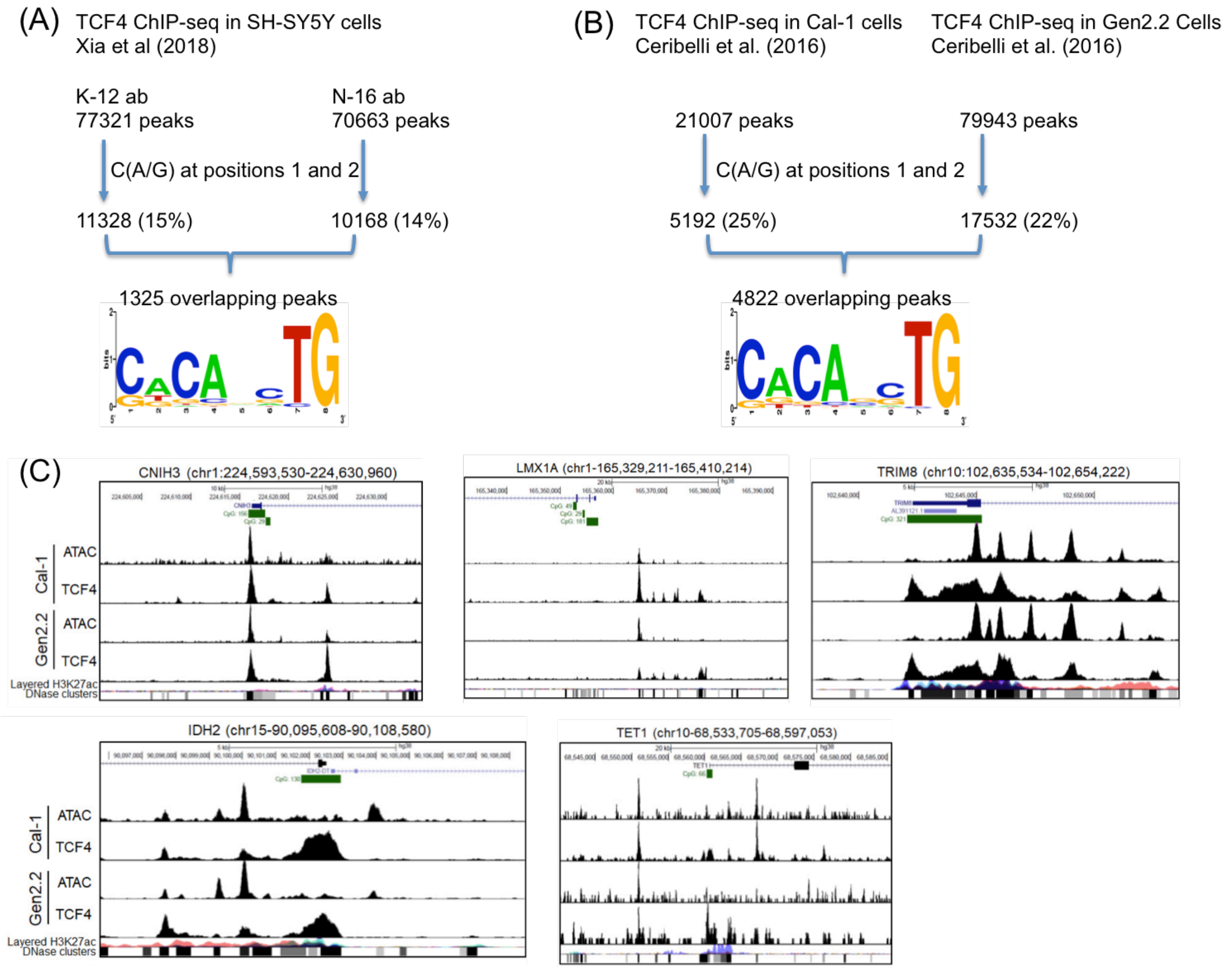
Summary of K_D values measured by ITC at [dsDNA]=3.3 μM and [TCF4]monomer=77.7 μM					
5′-AXGCACGTGXGT-3′ 3′-TGXGTGCACGXA-5′	X=C	X=M	X=H	X=F	X=5caC
K_D (μM)	0.37±0.10	0.18±0.03	0.49±0.12	0.18±0.04	0.027±0.009
N (sites)	2.18±0.06	2.28±0.03	1.83±0.05	1.63±0.03	1.93±0.02
ΔG (kcal/mol)	-8.78	-9.22	-8.61	-9.21	-10.3
ΔH (kcal/mol)	-10.1±0.5	-9.8±0.2	-9.1±0.5	-9.6±0.3	-9.5±0.3
-TΔS (kcal/mol)	1.33	0.53	0.52	0.34	-0.83

$\Delta G = \Delta H - T\Delta S$

Supplementary Figure S2. Related to Table 1E

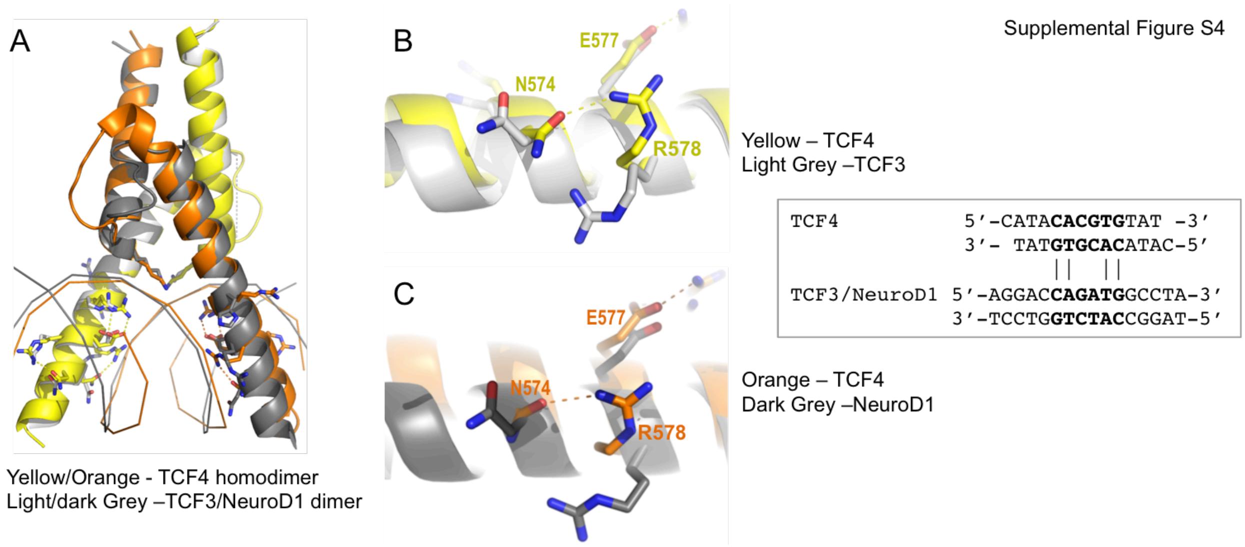
The ITC measurements of K_D values by various cytosine modifications of oligonucleotides were carried out under the conditions that the TCF4 protein ([monomer] = 77.7 μM) was titrated into the sample cell ([dsDNA] = 3.3 μM). (**A**) The raw data of an exothermic reaction releases heat that gives negative peaks. (**B**) The peaks are integrated as a function of molar ratio of [TCF4]/[DNA]. (**C**) The binding thermodynamic parameters (free energy ΔG, binding enthalpy ΔH and entropy -TΔS) are plotted for each modification. (**D**) The corresponding table summarizes the ITC parameters including equilibrium dissociation constant (K_D) and stoichiometry (N).

Supplemental Figure S3



Supplementary Figure S3. Examples of TCF4 binding sites containing 5' C(A/G) immediately adjacent to the E-box. The numbers of TCF4 binding sites (FDR/qvalue <0.05) were extracted from published ChIP-seq datasets. (**A**) Xia et al. used two different antibodies in neuroblastoma cells (SH-SY5Y): K-12 (GSE112704) and N-16 (GSE112704). Approximate 14-15% sites contain C(A/G) at positions 1 and 2. However, only 1325 (~8%) of these binding sites overlap between the two datasets, which were utilized to yield the binding motif. **(B)** Ceribelli et al. used two blastic plasmacytoid dendritic cell neoplasm cells: Cal-1 (ChIP-seq: GSM1975018 and ATAC-seq: GSM2243033) and Gen2.2 (ChIP-seq: GSM1975020 and ATAC-seq: GSM2243034). Approximately 22-25% sites contain C(A/G) at positions 1 and 2 and 4822 of these binding sites overlap between the two datasets, which we used to derive the binding motif. (**C**) Examples of genes associated with the TCF4 binding site containing 5' C(A/G) immediately adjacent to the E-box.

Supplemental Figure S4



Yellow/Orange - TCF4 homodimer
Light/dark Grey –TCF3/NeuroD1 dimer

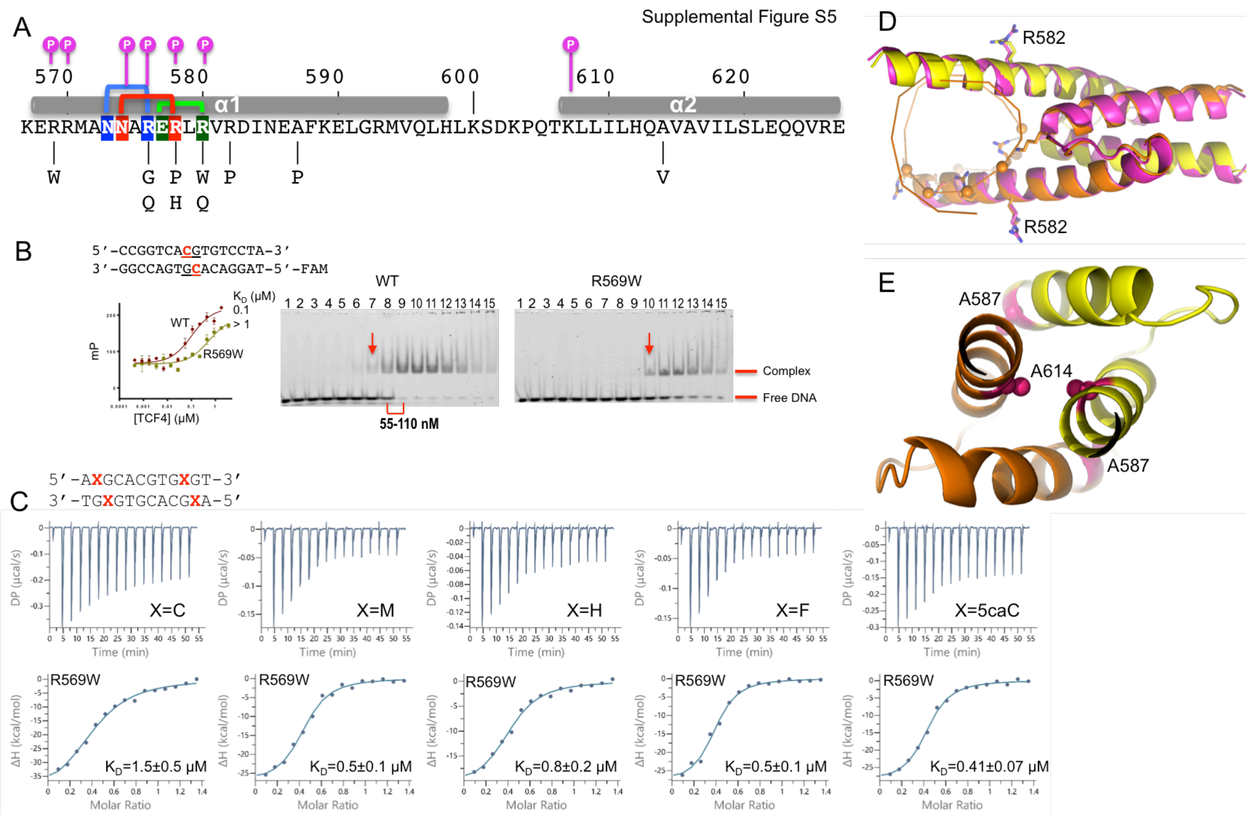
Yellow – TCF4
Light Grey –TCF3

Orange – TCF4
Dark Grey –NeuroD1

```
TCF4            5'-CATACACGTGTAT -3'
                3'- TATGTGCACATAC-5'
                      ||   ||
TCF3/NeuroD1 5'-AGGACCAGATGGCCTA-3'
             3'-TCCTGGTCTACCGGAT-5'
```

Supplementary Figure S4. Comparison between TCF4 and TCF3

(A) Superimposition of TCF4 homodimer (yellow and orange) and TCF3/NeuroD1 heterodimer (dark and light grey). (**B-C**) The N475•••R578 interaction in TCF4 is broken in TCF3 (panel C) and NeuroD1 (panel D) when in complex with asymmetric central dinucleotide. The single H-bond between G1 and R578 in TCF4 (Figure 5C) was absent in the TCF3-NeuroD1 heterodimer bound with an asymmetric sequence (5'-CATCTG-3') (PDB 2QL2).



Supplementary Figure S5. Mutant R569W

(A) Pitt-Hopkins mutations in TCF4 bHLH are indicated below the sequence, along with residues for making DNA phosphate contacts (white letter P in magenta background above the sequence). Three pairs of intra-molecular interactions exist in the major groove of DNA: N573•••R576 (blue), N574•••R578 (red) and E577•••R580 (green). (**B**) One Pitt-Hopkins associated mutant (R569W) diminished DNA binding as measured by FP and EMSA. (**C**) ITC measurement using DNA (40 μM) in syringe and R569W mutant (6 μM) in sample cell (20 mM Tris, pH 7.5 and 137 mM NaCl, 5% glycerol). (**D**) R582 is not engaged in DNA binding in TCF4-DNA cognate complex. (**E**) An A614V variant is likely to interfere with the dimer interaction. The L587P, located on the outer surface of helix 1, has the potential to alter helix conformation and the dimer interaction.

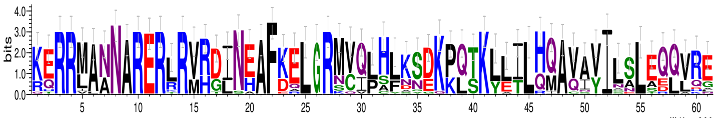
🟪 = Substitutions relative to *H. sapiens* TCF4; 🟨 = Residues changed in Pitt-Hopkins Syndrome

Class in Vertebrata	Species	TCF4 bHLH Region	GenBank #
Mammalia	*Homo sapiens*	KERRMANNARERLRVRDINEAFKELGRMVQLHLKSDKPQTKLLILHQAVAVILSLEQQVRE	CBY80191
Aves	*Gallus gallus*	KERRMANNARERLRVRDINEAFKELGRMVQLHLKSDKPQTKLLILHQAVAVILSLEQQVRE	Q90683
Reptilia	*Anolis carolinensis*	KERRMANNARERLRVRDINEAFKELGRMVQLHLKSDKPQTKLLILHQAVAVILSLEQQVRE	XP_016850729
Amphibia	*Xenopus tropicalis*	KERRMANNARERLRVRDINEAFKELGRMVQLHLKSDKPQTKLLILHQAVAVILSLEQQVRE	NP_001096226
Osteichthyes	*Lepisosteus oculatus*	KERRMANNARERLRVRDINEAFKELGRMVQLHLKSDKPQTKLLILHQAVAVILSLEQQVRE	XP_015217707
Chondrichthyes	*Rhincodon typus*	RERRMANNARERLRVRDINEAFKELGRMVQLHLKSDKPQTKLLILHQAVAVILSLEQQVRE	XP_020365982
Agnatha	*Lethenteron camtschaticum*		(none found)

Class in Vertebrata	Species	TCF12 bHLH Region	GenBank #
Mammalia	*Homo sapiens*	KERRMANNARERLRVRDINEAFKELGRMCQLHLKSEKPQTKLLILHQAVAVILSLEQQVRE	XP_011520261
Aves	*Gallus gallus*	KERRMANNARERLRVRDINEAFKELGRMCQLHLKSEKPQTKLLILHQAVAVILSLEQQVRE	XP_015134250
Reptilia	*Anolis carolinensis*	RERRMANNARERLRVRDINEAFKELGRMCQLHLKSEKPQTKLLILHQAVAVILSLEQQVRE	XP_016853254
Amphibia	*Xenopus tropicalis*	KERRMANNARERLRVRDINEAFKELGRMCQLHLKSEKPQTKLLILHQAVAVILNLEQQVRE	XP_002940299
Osteichthyes	*Lepisosteus oculatus*	RERRMANNARERLRVRDINEAFKELGRMCQLHLKSEKPQTKLLILHQAVAVILSLEQQVRE	XP_015198688
Chondrichthyes	*Rhincodon typus*	RERRVANNARERLRVRDINEAFKELGRMCQLHLNSDKPQTKLLILHQAVSVILNLEQQVRE	XP_020369765
Agnatha	*Lethenteron camtschaticum*		(none found)

Class in Vertebrata	Species	Atonal bHLH Region	GenBank #
Mammalia	*Homo sapiens*	ARRRLAANARERRRMQGLNTAFDRL-RRVVPQWGQDKKLSKYETLQMALSYIMALTRILAE	NP_660161
Aves	*Gallus gallus*	KQRRLAANARERRRMHGLNHAFDQL-RNVIPSFNNDKKLSKYETLQMAQIYISALAELLHG	XP_004941187
Reptilia	*Anolis carolinensis*	QTRRLLANARERTRVHTISAAFEAL-RKQVPCYSYGQKLSKLAILRIACNYILSLARLADL	XP_008117724
Amphibia	*Xenopus tropicalis*	KQRRLAANARERRRMHGLNHAFDQL-RNVIPSFNNDKKLSKYETLQMAQIYINALSDLLQA	XP_004911142
Osteichthyes	*Lepisosteus oculatus*	KQRRIAANARERRRMHGLNHAFDEL-RSVIPAFDNDKKLSKYETLQMAQIYINALSDLLQG	XP_006627369
Chondrichthyes	*Rhincodon typus*	KHRRLAANARERKRMHGLNHAFDEL-RSVIPAFDNDKKLSKYETLQMAQIYIAELTELLQN	XP_020365853
Agnatha	*Lethenteron camtschaticum*	KQRRLAANARERRRMHGLNHAFDRL-RNVIPSFAGDKKLSKYETLQMAQIYIGALAELLKG	AMN92150

Logo for all above sequences:



Supplementary Figure S6. Sequence conservation of bHLH of TCF4 related proteins

The basic helix-loop-helix (bHLH) regions of three related proteins are shown: TCF4 (the subject of this paper), TCF12 (a closely-related transcription factor), and Atonal. Sequences were identified using BlastP (National Center for Biotechnology Information). The species representing different Vertebrata classes are human, chicken, anole, clawed frog, spotted gar, whale shark, and arctic lamprey. Substitutions relative to human TCF4 are highlighted in purple; and residues that, when changed in human TCF4, are associated with Pitt-Hopkins Syndrome are highlighted in yellow. The logo at the bottom (from WebLogo; http://weblogo.threeplusone.com) shows conservation of residues among all three proteins from all seven species.

Supplementary Table S1. Summary of X-ray data collection from SERCAT (22-ID) at wavelength of 1 Å and refinement statistics in space group *P*1 (*)

DNA (5'-3') (3'-5')	CATACACGTGTAT TATGTGCACATAC	TTACACGTGTA ATGTGCACATT	AXGCACGTGXGT TGXGTGCACGXA (X=5caC)
PDB Code	6OD3	6OD4	6OD5
Number of crystals	1	1	2
Cell dimensions (Å)	41.18, 58.95, 62.72	36.61, 43.60, 43.56	44.68, 44.76, 54.64
α, β, γ (°)	104.6, 90.3, 94.9	97.2, 102.6, 102.5	78.6, 78.9, 79.3
Resolution (Å)	37.90-1.49 (1.54-1.49)	41.82-1.70 (1.75-1.70)	36.66-2.05 (2.12-2.05)
[a] R_{merge}	0.056 (0.730)	0.103 (0.711)	0.063 (0.554)
R_{pim}	0.033 (0.471)	0.041 (0.418)	0.038 (0.412)
$CC_{1/2}$, CC	(0.627, 0.878)	(0.522, 0.828)	(0.361, 0.729)
[b] <I/σI>	21.6 (1.6)	16.4 (3.7)	5.4 (2.0)
Completeness (%)	84.1 (70.6)	95.3 (83.1)	96.5 (82.8)
Redundancy	3.6 (2.6)	6.2 (3.2)	3.6 (2.2)
Observed reflections	320,996	160,484	88,261
Unique reflections	88,521 (7433)	26,021 (2269)	24,281 (2094)
Refinement			
Resolution (Å)	1.49	1.69	2.05
No. reflections	88,360	25,959	24,247
[c] R_{work} / [d] R_{free}	0.221 / 0.237	0.222 / 0.267	0.233 / 0.279
No. Atoms			
Protein	3718	1755	1957
DNA	1032	407	962
Solvent	410	44	163
B Factors (Å^2)			
Protein	40.2	46.0	30.6
DNA	31.2	55.2	33.3
Solvent	42.4	46.6	35.4
R.m.s. deviations			
Bond lengths (Å)	0.002	0.002	0.003
Bond angles (˚)	0.4	0.4	0.6

* Values in parenthesis correspond to highest resolution shell.

[a] R_{merge}=Σ|I-<I>|/ΣI, where I is the observed intensity and <I> is the averaged intensity from multiple observations.

[b] <I/σI> = averaged ratio of the intensity (I) to the error of the intensity (σI).

[c] R_{work}=Σ|Fo-Fc|/Σ|Fo|, where Fo and Fc are the observed and calculated structure factors, respectively.

[d] R_{free} was calculated using a randomly chosen subset (5%) of the reflections not used in refinement.