Identification of CTCF as a master regulator of the clustered protocadherin genes

Michal Golan-Mashiach¹, Moshe Grunspan¹, Rafi Emmanuel¹, Liron Gibbs-Bar¹, Rivka Dikstein¹,* and Ehud Shapiro¹,2,*

¹Department of Biological Chemistry and ²Department of Computer Science and Applied Math, Weizmann Institute of Science, Rehovot 76100, Israel

Received July 27, 2011; Revised December 3, 2011; Accepted December 6, 2011

ABSTRACT

The brain is a large and complex network of neurons. Specific neuronal connectivity is thought to be based on the combinatorial expression of the 52 protocadherins (Pcdh) membrane adhesion proteins, whereby each neuron expresses only a specific subset. Pcdh genes are arranged in tandem, in a cluster of three families: Pcdhα, Pcdhβ and Pcdhγ. The expression of each Pcdh gene is regulated by a promoter that has a regulatory conserved sequence element (CSE), common to all 52 genes. The mechanism and factors controlling individual Pcdh gene expression are currently unknown. Here we show that the promoter of each Pcdh gene contains a gene-specific conserved control region, termed specific sequence element (SSE), located adjacent and upstream to the CSE and activates transcription together with the CSE. We purified the complex that specifically binds the SSE-CSE region and identified the CCTC binding-factor (CTCF) as a key molecule that binds and activates Pcdh promoters. Our findings point to CTCF as a factor essential for Pcdh expression and probably governing neuronal connectivity.

INTRODUCTION

The brain consists of a large and complex organism-specific network of neurons. A fundamental question in neurobiology is where and how the structure of this network is encoded in each organism’s genome, i.e. how individual neurons acquire unique identities that enable them to create highly-specific synaptic connections, leading to the formation of an organism-specific network. There are several well-established examples for such mechanisms: the olfactory system, in which individual olfactory neurons express only 1 out of 1300 olfactory receptor genes, and establish connections based on the receptor expressed (1), and the Down-Syndrome-Cell-Adhesion-Molecules (Dscam) that regulate neural circuit formation in Drosophila (2–4). In mammals, it was hypothesized that the precise patterns of neuronal connectivity are largely determined by neuronal membrane molecules called protocadherins (Pcdh), which promote specific inter-neuron connections. These molecules are encoded by the clustered Protocadherin (Pcdh) genes (5–11) that represent the largest subgroup in the cadherin superfamily (12,13). The Pcdh genes are present in all known vertebrate genomes, including mammals, chicken, zebrafish, fugu, and coelacanth (13) and are highly conserved in mammals. The Pcdh genes were shown to be highly expressed during neural development, creating Pcdh proteins that are concentrated in the synaptic region. As the brain matures, the expression level of the Pcdh genes decreases (10,12,14–17). Gene-knockout studies have demonstrated that Pcdh gene products play a crucial role in proper axonal projection, synaptic formation and neuronal survival (18,19).

Pcdh genes are located on human chromosome 5 (13) and on mouse chromosome 18 (20,21). In each neuron, Pcdh genes are expressed monoaletically, each allele is independently regulated, i.e. one variable exon is expressed from the paternal chromosome and another variable exon from the maternal chromosome, creating a combinatorial expression at the cell level (22,23). There are 52 tandem-arranged Pcdh genes in human that are divided into three families: Pcdhα (15 genes), Pcdhβ (16 genes), and Pcdhγ (21 genes), which are further subdivided into Pcdhαa and Pcdhαb (21).

In the Pcdhα and Pcdhγ families, each gene has a specific variable exon (24) that is linked to three constant exons (a structure similar to that of immunoglobulins and T-cell-receptors). In the Pcdhβ family, each gene includes a variable exon only. Linking of a specific variable exon with the three constant exons in
the Pedhα and Pedhγ genes is done by alternative splicing (as opposed to gene recombination used in immunoglobulin genes). Prior to splicing, each precursor mRNA transcribed by the Pedhα and Pedhγ genes, is of high molecular weight, since it includes all downstream variable exons (25). During splicing, only the 5′-most variable exon is spliced to the first constant exon to generate functional mRNAs (13,21,26).

Assuming that each cluster expresses, at the most, one gene from each allele, this unusual Pedh expression may provide sufficient diversity to represent at least 1.5 million unique individual cell labeling (see materials and methods for the calculation). Each Pedh variable exon is preceded by a distinct promoter, and all promoters contain a similar highly conserved core motif of ~22 bp, the conserved sequence element (CSE) (26,27). In addition, long-range cis-regulatory DNA elements in the Pedhα gene cluster, HS5-1 and HS7, were identified and found to possess enhancer activity in reporter assays (28). Interestingly, both elements are conserved among vertebrates and also include a CSE. The Pedhα genes are likely to be regulated by methylation as the transcription of specific Pedhα genes was found to be significantly correlated with the methylation state of the first exon. On the other hand, mosaic or mixed methylation states of the CSE in the promoters were associated with both active and inactive transcription (29). Presently, the mechanisms underlying promoter choice and promoter activity are largely unknown.

In the present study, we investigated the mechanism underlying Pedh gene transcription. Using bioinformatics methods, we have identified a sequence element located near the CSE that is highly conserved among mammals but, unlike the Common Sequence Element, is unique to each of the 52 Pedh genes. We termed the identified element specific sequence element (SSE). The 20-bp-long SSE is essential for transcription and can activate transcription only in the presence of the CSE. We have purified the complex that binds the SSE–CSE region and identified CCTC-binding factor (CTCF) as a factor that binds the promoters of Pedh genes through the common CSE and the SSE elements. Remarkably in the context of Pedh genes CTCF plays an essential positive role in transcription. Our findings point to CTCF as a factor that is essential for Pedh expression and probably for the control of neuronal connectivity.

MATERIALS AND METHODS

Bioinformatics analysis of the Pedh promoters

Regulatory elements such as enhancers and locus control regions are highly conserved among different mammalian species. To identify regions containing putative DNA elements that regulate Pedh expression, we compared the genomic DNA upstream to the first exon of each gene in the three Pedh families (α, β and γ). Human, chimp, mouse, rat and dog promoter regions from −1000 relative to the TSS were retrieved from UCSC Genome Browser (http://genome.ucsc.edu/). This analysis revealed new conserved sequences located immediately upstream of the previously identified CSE. These regions are more conserved among species (orthologs) than among family members within the same species (paralogs) (Supplementary Figure S1).

For calculation of combinational Pedh diversity, we used binomial coefficients: the number of ways to choose K elements from N is \( \binom{N}{k} = \frac{N!}{k!(N-k)!} \) for 0 ≤ k ≤ n.

The calculation is as follows: the number of ways to choose two variable exons from 13 (for Pedhα), 13!/11! × 2!, multiplied by choose 2 from 15 (for Pedhβ), 15!/13! × 2!, multiplied by choose 2 from 19 (for Pedhγ), 19!/17! × 2!, is equal to 1.5 million unique labels. This calculation does not take into account that the Pedhα and Pedhγ proteins also form oligomers (30), which can further increase the molecular diversity at the cell surface.

Plasmid construction

The promoter regions of the α6 and α3 genes (−217 and −219 relative to translation start site, respectively) were cloned by genomic PCR into pGL3-Basic (Promega) using the HindIII site. The promoters were associated with both active and inactive transcription state of the first exon. On the other hand, mosaic or mixed methylation states of the CSE in the promoters were associated with both active and inactive transcription (29).

For determining the effect of CTCF depletion on the activity of the reporter gene, SH-SY5Y cells were grown on 10-well plates and transfected with CTCF siRNA (Origene, cat # SR307273), or Scrambled siRNA-Scr (Origene, cat # SR30004) as a negative control at a concentration of 10 nM. Twenty-four hours later, the cells were split to 12-well plates. After an additional 72 h (96 h after siRNA transfection) the cells were transfected with 1.5 μg luciferase reporter plasmid, 50 ng CMV-GFP, 25 ng CMV-GFP, 50 ng Rous sarcoma virus (RSV) promoter-driven Renilla luciferase reporter plasmid, At 24 h after transfection, cells were harvested and their luciferase and Renilla luciferase activities were measured. The mean and the SD values were determined for each construct based on four independent transfections.

For reporter assays, subconfluent cells were transfected in a 24-well plate using 750 ng luciferase reporter plasmid, 25 ng CMV-GFP, 50 ng Rous sarcoma virus (RSV) promoter-driven Renilla luciferase reporter plasmid. At 24 h after transfection, cells were harvested and their luciferase and Renilla luciferase activities were measured. The mean and the SD values were determined for each construct based on four independent transfections.

Nucleic Acids Research, 2012, Vol. 40, No. 8 3379
100 ng Rous sarcoma virus (RSV) promoter-driven *Renilla* luciferase reporter plasmid. At 24 h after second transfection, cells were harvested and their luciferase and *Renilla* luciferase activities were measured. The mean and the SD values were determined for each construct based on four independent transfections.

For determining the effect of CTCF depletion on the activity of the endogenous *Pcdh* mRNA levels, HEC1-B cells were grown on 6-well plate and transfected with 10 nM CTCF siRNA (Origene, cat # SR307273), or Scrambled siRNA-Scr (Origene, cat # SR30004) as a negative control. One hundred and twenty hours later RNA was extracted using Tri-reagent (MRC Inc.).

cDNAs were synthesized from 1 μg of total RNA in a 20 μl reaction volume using the SuperScript III Reverse Transcriptase (Invitrogen) and random primers as per the manufacturer's instructions. Quantitative real time PCRs were performed in duplicate using the Applied Biosystems 7300 Real-Time PCR System (Applied Biosystems). Independent PCRs were performed using the same cDNA for genes of interest and the GAPDH gene which serves as an internal control, using the SYBR® Green PCR Master Mix (Applied Biosystems). Gene-specific primers were designed for the genes of interest and for the GAPDH gene using Primer Express® software (Applied Biosystems). The sequences of the primer pairs are listed in Supplementary Data S3.

**Electrophoretic mobility shift assay and ‘super-shift’ assay**

DNA oligonucleotides containing the α6SSE–CSE or α3SSE–CSE or α12SSE–CSE or SP1 sequences (Supplementary Data S2, oligonucleotides 1–2, 10 and 18) were fluorescently labeled on the 5'-end with Cy5, HEX, or FAM (Integrated DNA Technologies, Inc). The oligonucleotides were annealed and then ligated with the same cDNA for genes of interest and the GAPDH gene using Primer Express® software (Integrated DNA Technologies). The sequences of the primer pairs are listed in Supplementary Data S3, primers 1–5.

**Fractionation of nuclear extract by Macro-prep high S cation exchange column**

SH-SY5Y nuclear extract prepared as described previously was incubated with washed Macro-prep high S cation exchange support beads (Bio-Rad) at 4°C for 30 min. Flow through fractions where collected after separation from the beads using centrifugation at 500 G for 4 min. Elution with 200, 400 and 800 mM KCl buffers containing 25 mM HEPES (pH 7.9), 1 mM EDTA, 1 mM dithiothreitol and 10% glycerol was performed using centrifugation at 500 G for 4 min. Dialysis over night was performed on eluted fractions using 10 KDa dialysis bag (Thermo scientific) in dialysis buffer described previously. A fraction of all the eluted samples was analyzed on SDS–PAGE gradient acryl amid gel 4–12.5% (GEBA) with 1× Tris–Glycine 0.1% SDS buffer at 130 V for 1 h.

**Affinity purification and western blot**

Affinity magnetic FG plain beads were prepared as follows: oligonucleotides of the α6 and α6_mut sequences, with phosphorylated GGGG or CCCC at the 5'-end of the forward and reverse complement, respectively, were synthesized (Integrated DNA Technologies, Inc) (Supplementary Data S2, oligonucleotides 22–23). The oligonucleotides were annealed and then ligated using T4 ligase 5 U/μl (Fermentas) using the manufacturer ligation buffer for 16 h in 4°C to generate a concatamer of 5–15 repeats. The muted sequence has the same sticky ends. The ligated DNA was then phenol–chloroform extracted and desalted using Nick column (GE healthcare) equilibrated with water. The ligated samples were run on agarose electrophoresis gel to verify that the ligation process has succeeded (Supplementary Figure S5). The ligated DNA was then immobilized to 10 mg magnetic FG plain beads (Tamagawa seiki Co. Ltd) by incubation at 50°C for 24 h. The immobilized beads where then washed with 2.5 M KCl and incubated with 1 M ethanalamine (pH 8) over night for masking. Prior to purification the immobilized beads were washed three times using a magnetic stand in washing buffer containing 25 mM HEPES (pH 7.9), 100 mM KCl, 1 mM EDTA, 1 mM dithiothreitol and 10% glycerol. Crude nuclear extract or elution fraction collected from cation fractionation was incubated with beads at 4°C for 4 h using rotator in binding buffer containing 2 μg of poly(dI–dC) (Sigma), 1 μg of poly(dA–dT) (Sigma) consisting of 25 mM HEPES (pH 7.9), 50 mM KCl, 1 mM DTT and 10% glycerol. Beads were then washed four times using a magnetic stand with washing buffer previously described. Bound proteins were then eluted from the beads using a magnetic stand with elution buffer at room temperature for 30 min consisting of, 25 mM HEPES (pH 7.9), 1 M KCl, 1 mM EDTA, 1 mM dithiothreitol and 10% glycerol. Dialysis over night was performed on a fraction of the sample using dialysis bag (Thermo scientific) in dialysis buffer described previously. A fraction of the sample was sent to Mass-spec analysis for protein identification. Tandem mass spectrometry (MS/MS) analysis coupled with liquid chromatography...
(LC) was carried out by the Smoler Proteomics Center (Faculty of Biology, Technion, Israeli Institute of Technology, Israel). Top candidates are those proteins that were identified by mass spectrometry analysis following DNA affinity purification only from the wild-type SSE–CSE but not detected at the elution of the mutated SSE–CSE sequences in three repeats.

For western blot, a fraction of the sample was analyzed on SDS–PAGE gradient acryl amid gel 4–12.5% (GEBA) with 1× Tris–Glycine 0.1% SDS buffer at 130 V for 1 h, transferred to nitrocellulose membranes blots, and probed with anti-CTCF (Abcam ab70303) and anti-Histone H3 (Abcam ab1791) antibodies. Secondary antibodies consisted of goat anti-rabbit conjugated to IRDye800 or to IRDye680 (LI-COR Biosciences), were used and then the membrane was scanned for infrared signal using the Odyssey Imaging System (LI-COR Biosciences).

Production of CTCF Proteins in vitro

Full-length human CTCF (pET-7.1) and the 11 ZF CTCF-binding domain (pET-11ZF) which were verified in several papers (33,34), were kindly donated by V. Lobanenkov (National Institute of Allergy and Infectious Diseases, National Institutes of Health, Rockville, MD, USA). These plasmids along with Luciferase T7 Control DNA, which serves as a positive control, were synthesized using the TnT® T7 Quick Coupled Transcription/Translation System (Promega). The in vitro synthesized proteins were labeled with a fluorescently labeled lysine amino acid using the FluoroTect GreenLys labeling system according to the manufacturer’s instructions (Promega). We verified the identity of the plasmids using restriction enzyme analysis and SDS–PAGE analysis which verified that the synthesized proteins have the correct molecular mass (Figure 6A).

RESULTS

Identification of a novel cis-element in the promoters of the Pcdh genes

In order to identify the regulatory elements that control Pcdh expression, we first looked for additional segments in the Pcdh cluster that are conserved among mammalian species, namely: chimp, mouse, rat, dog and human. The 1000 bp upstream of each of the 52 Pcdh V-exons from the above species were aligned. This analysis revealed novel highly conserved regions located immediately upstream of the previously defined conserved sequence element, CSE (16,35,36) (Figure 1). Interestingly, unlike the CSE which is common to the promoters of all families, these conserved regions, although they may have common regions, are unique to each gene (Supplementary Figure S1). We therefore termed these regions, Specific Sequence Elements (SSE). All 52 occurrences of the CSE and the 52 unique SSEs are highly conserved in mammals (Supplementary Figure S1). The SSEs we analyze consist of α(1 to 13)SSE, β(1 to 16)SSE, γα(1 to 12)SSE and γb(1 to 7)SSE. The PcdhaC1, αC2, PcdhyC3, γC4, and γC5 were omitted from this analysis.

The αSSE is essential for Pcdh transcriptional activity

To investigate the transcriptional control of the Pcdh gene clusters we chose to focus on the Pcdha family. We first examined the expression of all the Pcdha genes in HEC-1B, 293T and SH-SY5Y cell lines by RT–PCR and found that each of these cell lines expresses at least one Pcdha gene (Figure 2A). While the CSE has been previously shown to be important for the transcription of Pcdh genes (16,35,36) we wished to examine the function of the newly discovered αSSE in the Pcdha family, and for this purpose Pcdha6 was selected. The α6 promoter containing the α6SSE and the CSE was cloned in front of a luciferase

Figure 1. Schematic representation of the Pcdh gene cluster promoters. Genomic organization of the Pcdh genes, with tandem variable region exons (blue), promoter regions (turquoise) and constant region exons (red). Each promoter region contains an SSE followed by a CSE, which is common to all genes of the family. V, variable region; C, constant region.
reporter gene (Figure 2B). Upon transfection into HEC-1B, 293T and SH-SY5Y cell lines, this promoter fragment displayed a high transcriptional activity compared to the promoter-less (basic) reporter gene in all these cells (Figure 2C, WT columns). When the α6SSE or α6SSE plus CSE were deleted (Δα6SSE or Δα6SSE-ΔCSE, respectively) the transcriptional activity was severely diminished in all the cell lines (Figure 2C). Likewise, point mutations in the CSE (CSEmut) also diminished transcription (Figure 2C), in agreement with previous reports (16,35,36). We also examined the SSE function of the Pcdh/C11 gene promoter (α3SSE) and found it to be essential for α3 promoter activity (Figure 2C, fourth panel). These findings confirm that CSE is crucial but not sufficient for Pcdh expression and also requires in addition the SSE to drive the Pcdh transcription.

To investigate α6SSE further, α6SSE mutants were created by site-directed mutagenesis (Figure 3A, Mut1–Mut4). The α6SSE Mut2 construct reduced the transcriptional activity to the level of the promoter-less (basic) construct, Mut3–Mut4 promoters caused reduction of 70% while Mut1 did not have a significant effect on the transcription level (Figure 3A).

To determine whether the distance between the SSE and the CSE is important for their function we increased the spacing between the SSE and the CSE sites by 5, 10 and 15 bp, which corresponded to 0.5, 1 and 1.5 helical turns (5nLinker, 10nLinker and 5nLinker). The results revealed that the transcriptional activity was reduced to the level of promoter-less construct in all the examined cell lines, regardless of the size of the spacer that was introduced (Figure 3B). These findings suggest that the specific location of the SSE relative to the CSE is crucial for their promoter activity.

Identification of a specific DNA-binding complex that is αSSE and CSE dependent

To investigate further the mechanism underlying αSSE–CSE activity we analyzed the proteins that interact with
the αSSE–CSE region. A fluorescently labeled DNA fragment comprising the αSSE–CSE sequence was incubated with a nuclear extract prepared from HEC-1B cells and then subjected to electrophoresis mobility shift assay (EMSA). Two major complexes were formed between the DNA and the extract (Figure 4A). The complexes were competed out by excess (50-fold) of unlabeled αSSE–CSE DNA (Figure 4A, lane 2) but not by excess of the non-relevant Sp1 sequence DNA (Figure 4A, lane 3).

Figure 3. Sequence requirements for the function of αSSE–CSE as transcriptional elements. (A) Successive blocks within αSSE (underlined) in the αSSE–CSE promoter were mutated (Mut1-Mut4). The wild-type and mutated constructs were transfected into HEC-1B, 293T and SH-SY5Y cell lines together with RSV-Renilla that serves as internal control. Twenty-four hours post-transfection firefly and Renilla luciferase activities were measured. The normalized results are the mean of at least four independent experiments (±SD). (B) Linkers of 5, 10 and 15 bp (5nLinker, 10nLinker and 5nLinker) were introduced between αS6SSE and CSE and their effect was analyzed as described earlier.
To examine the importance of the a6SSE–CSE region, competition assays with excess of various types of unlabeled DNA were performed: (i) DNA with mutations in either a6SSE, CSE or both (ii) DNA with spacing between the a6SSE and CSE regions. We found that the complex was not competed out by excess DNA containing mutations in either the CSE (Figure 4B, lane 4), a6SSE (Figure 4B, lane 5), both (Figure 4B, lane 6) or mut2 (Figure 4B, lane 7).
z6SSE (Figure 4B, lane 7). In addition, a 5-bp linker between the z6SSE and CSE also failed to compete (Figure 4B, lane 8) indicating that the spacing between the elements is important for binding. These results are consistent with the effect of the mutations and the linker on promoter activity (Figure 3B) and suggest that the two elements are critical for the DNA binding complex.

A recent study reported that the CSE shows mosaic methylation or hypo-methylation, regardless of transcription level (29). Our findings also show that CpG methylation of the zSSE–CSE regions does not affect complex formation (Supplementary Figure S3).

Next, we carried out competition assays with a similar DNA sequence derived from the other promoters of the Pcdh cluster (with z1SS–CSE, z2SS–CSE, . . . , z13SS–CSE) and of the Pdhy cluster, γA2SS–CSE. Remarkably, the complex was not competed out by excess of most of the Pedha SSE–CSE and not by excess of the Pcdh γA2SS–CSE (Figure 4C). We similarly analyzed the complex that binds to the z3SS–CSE (Figure 4D, left panel) and the z12SS–CSE (Figure 4D, right panel). A competition assay was performed, in which the z3SS–CSE or z12SS–CSE labeled DNA probes were competed out by excess of unlabelled DNA from parallel regions of other Pcdh genes (Figure 4D). The z3SS–CSE and z12SS–CSE complexes failed to be competed out by SSE–CSE regions corresponding to other genes in the cluster. The results suggest the existence of a unique component for each SSE that confers specificity of the complex that binds to each promoter of the Pedha.

**Purification of the zSSE–CSE complex and its identification by mass spectrometry**

To further characterize zSSE–CSE binding proteins, we developed a purification scheme composed of two chromatographic steps (Figure 5A). First, SY5Y nuclear extract was loaded onto Macro-prep high S cation exchange column and proteins were eluted by increased salt concentrations. The fractions where analyzed by EMSA and the z6SSE–CSE specific complex was found in the 800 mM KCl fraction (Figure 5B, lane 2), resulting in ~3-fold enrichment. Next, the 800 mM fraction was subjected to DNA affinity chromatography using magnetic nanoparticles (37). The affinity matrix was generated by chemical cross-linking immobilization to the magnetic nanoparticles the z6SSE–CSE or the z6SSE–CSE_mut DNAs that wereimerized. The z6SSE–CSE_mut DNA was mutated in both the z6SSE and the CSE elements and served as negative control. After incubation of the 800 mM enriched fraction with the DNA beads, unbound and eluted fractions were analyzed by EMSA using z6SSE–CSE DNA (Figure 5B). It was apparent that the z6SSE–CSE was efficiently depleted from the unbound fraction of the z6SSE–CSE DNA containing beads (Figure 5B, lane 3). In contrast, in the unbound fraction of the z6SSE–CSE_mut the specific complex was still present (Figure 5B, lane 4). Consistently, the elution of the z6SSE–CSE DNA affinity contained the z6SSE–CSE complex whereas the complex was completely absent from the elution of the mutated DNA beads (Figure 5B, lanes 5 and 6). This purification step resulted in approximately a 13-fold enrichment. To identify the proteins that specifically bind to z6SSE–CSE, the affinity eluted fractions of both z6SSE–CSE and z6SSE–CSE_mut DNA were subjected to comparative mass-spectrometry. The top significant candidates from mass-spectrometry analysis of proteins that specifically bind to the z6SSE–CSE are depicted in Table 1.

To test which of the proteins identified by the mass-spectrometry binds specifically to z6SSE–CSE, we examined whether antibodies against these factors interfere or ‘super-shift’ the z6SSE–CSE complex in EMSA. Three different anti-CTCF antibodies were added to the reaction mix. The first anti-CTCF antibody (ab70303) eliminated the specific transcription complex (Supplementary Figure S4) while the two other anti-CTCF antibodies (ab37477, ab37478) retarded the migration of the specific transcription complex and a slower migration was formed instead of the original one (Figure 5C, lanes 2 and 3). Interestingly this was the case also for another, non-specific complex which was super-shifted with anti-BUB3 antibody (Figure 5C, lane 5). In addition we used an antibody for YY1, a transcription factor that appeared in the mass-spectrometry results in both the z6SSE–CSE and the control z6SSE–CSE_mut DNA. The addition of anti YY1 to the reaction mix retarded another complex that binds to the z6SSE–CSE fluoresently labeled probe (Figure 5C, lane 4), which was not specific to these elements. Inspection of the sequence z6SSE–CSE DNA that was used as a probe revealed the presence of a perfect YY1 binding site at the 3′-end of the fragment which was unaffected by the mutations. These results imply that there are several protein-DNA complexes formed between the nuclear extract and the z6SSE–CSE labeled DNA probe and provide validation that the purification scheme we utilized led to the identification of relevant proteins. Next, we examined the effect of the CTCF antibody on the complexes formed with a DNA fragment from other members of the Pedha family (z12SS–CSE and z3SS–CSE) or from another family (γA2SS–CSE). The same pattern, which eliminated the specific complex (ab70303), was observed with all these probes (Supplementary Figure S4). These findings are consistent with in vivo data as analysis of the recently released ChIP-seq data from UCSC genome-browser and from Handoko et al. (38) which revealed that CTCF was associated with the promoter of each of the Pedh genes in vivo.

To further validate that CTCF binds specifically to the SSE–CSE we performed western blot following affinity purification using anti-CTCF (Abcam ab70303). The CTCF was depleted from the unbound fraction of the WT but not the mutated DNA beads. Consistently, CTCF was exclusively eluted from the WT and not in the mutated affinity column (Figure 5D).

To gain further support that SSE–CSE is directly bound by CTCF, we used in vitro transcribed and translated full-length human CTCF and 11ZFs fragments (Figure 6A) in EMSA. A specific retarded complex was
observed using 11ZFs protein for DNA promoters of several members of the Pcdh family: α6SSE–CSE (Figure 6B, lane 8), α12SSE–CSE (Figure 6B, lane 1) and α3SSE–CSE (Figure 6B, lane 4). Recombinant luciferase protein prepared by the same in vitro translation reaction was used as a negative control for site-specific DNA-binding experiments (Figure 6B, lanes 2, 4, 13). No complex observed, however, with the full-length CTCF (Figure 6B, lane 7) possibly since the DNA binding domain of CTCF in the full-length context is masked or not properly folded. The specificity of the interaction was confirmed by a competition assays with excess of unlabeled DNA sequence of WT (Figure 6B, lane 9) and DNA with mutations in either CSE (Figure 6B, lane 10) α6SSE (Figure 6B, lane 11), or both (Figure 6B, lane 12). The in vitro data provides strong evidence for the novel interaction of CTCF with highly conserved CSE-SSE sequences present in the Pcdh gene promoters. These sequences were sufficient for function, for binding and for biochemical purification of CTCF.

**CTCF is essential for Pcdh expression**

To investigate further the role of CTCF in regulation of Pcdh expression, we tested whether knocking down the endogenous CTCF would affect Pcdh promoter activity. SH-SY5Y cells were transfected with siRNA against CTCF or a scrambled negative control. Seventy-two hours post-transfection Pcdhα6 promoter-luciferase reporter genes were transfected into these cells and
Table 1. Significant candidate proteins which specifically bind to the α6SSE–CSE identified by mass-spectrometry analysis

| Accession Number | Protein name                      | Gene symbol | Function                                                                 | Number of peptides |
|------------------|-----------------------------------|-------------|--------------------------------------------------------------------------|-------------------|
| 1                | Transcriptional repressor         | CTCF        | Transcription regulator and chromatin insulator                        | 4                 |
| 2                | Mitotic checkpoint protein        | BUB3        | Involved in spindle checkpoint function                                 | 3                 |
| 3                | Transcription elongation regulator | TCR1        | Regulates transcriptional elongation and pre-mRNA splicing              | 3                 |
| 4                | Thymocyte nuclear protein 1       | THYN1       | May be involved in the induction of apoptosis                            | 3                 |
| 5                | Negative elongation factor E      | NELFE       | Part of a complex which represses RNA polymerase II transcript elongation | 2                 |
| 6                | Zinc finger protein 691           | ZN316       | May be involved in transcriptional regulation                            | 3                 |
| 7                | Hepatoma-derived growth factor    | HDGF        | Acts as a transcriptional repressor                                      | 2                 |
| 8                | Transcription factor BTF3 homolog 4 | BT3L4      | Basic transcription factor—unknown                                      | 2                 |
| 9                | Nucleophosmin                     | NPM         | Involved in diverse cellular processes such as ribosome biogenesis, centrosome duplication, protein chaperoning, histone assembly, cell proliferation, and regulation of tumor suppressors TP53/p53 and ARF | 2                 |
| 10               | Zinc finger protein 768           | ZN768       | May be involved in transcriptional regulation                           | 2                 |
| 11               | Polyglutamine-binding protein     | PQBP1       | Involved with transcription activation                                  | 2                 |
| 12               | Zinc finger protein 787           | ZN787       | May be involved in transcriptional regulation                            | 2                 |
| 13               | LIM domain-binding protein 1      | LDB1        | Binds to the LIM domain of a wide variety of LIM domain-containing transcription factors | 3                 |
| 14               | Zinc finger protein 316           | ZN316       | May be involved in transcriptional regulation                            | 3                 |
| 15               | C-Myc-binding protein             | MYCBP       | Stimulates the activation of E box-dependent transcription by MYC      | 5                 |
| 16               | Ubiquitin-associated protein 2    | UBAP2       | The function of this protein has not been determined                    | 5                 |
| 17               | Uncharacterized protein C14orf43  | CN043       | unknown                                                                  | 6                 |

Figure 6. Recombinant CTCF binds specifically to several αSSE–CSE DNA sequences. (A) SDS-PAGE analysis of the full-length CTCF and 11ZF CTCF-binding domain proteins, which were synthesized in vitro from the pET-7.1 and pET-11ZF constructs. Luciferase T7 Control DNA no plasmid were used for positive and negative controls, respectively. The in vitro synthesized proteins are fluorescently labeled. Positions of the molecular mass protein markers (on the left) are indicated. The white arrows point to the positions of the in vitro synthesized proteins. (B) EMSAs using in vitro-translated luciferase, human CTCF full-length (FL), 11ZF or SH-SY5Y nuclear extract with αSSE–CSE sequences as probes. Lanes 1–3, α12SSS–CSE probe, lanes 4–6, α3SSS–CSE probe and lanes 7–14, α6SSS–CSE probe. The proteins used for binding and the competitor DNA are indicated on the top. The specific complexes between the probe and the recombinant proteins or the nuclear extract are indicated by arrows.
luciferase activity was determined 24 h later. Down-regulation of endogenous CTCF by siRNA-CTCF but not scrambled siRNA was confirmed by western blot analysis with anti-CTCF (ab70303) (Figure 7A). The results clearly show that CTCF siRNA but not scrambled siRNA severely diminished the luciferase activity directed by the WT Pcdh/C11 6 promoter by >3-fold (Figure 7A). This reduction in luciferase activity was specific to Pcdh because siRNA-CTCF did not affect the luciferase activity driven by the SV40 promoter (Figure 7A).

To gain further support for the involvement of CTCF in the transcriptional activity of endogenous Pcdh mRNA, we measured the mRNA levels of Pcdha6 and Pcdha12 in HEC1-B cells that selectively expressed these genes, following knock down of CTCF by siRNA (Figure 7B). In these cells CTCF protein level was decreased by 80% whereas its levels were unchanged by the scrambled siRNA. Down-regulation of CTCF caused similar reduction in the levels of the endogenous Pcdha6 and Pcdha12 mRNA levels. As a positive control, we selected the c-Myc gene, for which CTCF was known to act as a repressor (39), and found that c-Myc levels were up-regulated as expected. These findings strongly suggest that CTCF acts as a major activator of Pcdh genes.

**DISCUSSION**

In the present study we have identified a novel specific element (SSE) found in the promoters of all 52 Pcdh genes, positioned immediately upstream of each CSE. In contrast to the CSE, SSEs are rather unique sequences and much less conserved between paralogs of the same gene family. We have shown that Pcdh promoter activity was governed by the combined activity of the newly identified SSE and the CSE. Thus, while CSE is indeed crucial for transcription, it is definitely not sufficient. Using DNA binding assay we found a nuclear complex that specifically interacts with a SSE–CSE in vitro. Subsequently, we have isolated and identified CCTC-binding-factor (CTCF) as the protein that binds and specifically binds and regulates the SSE–CSE in each promoter region of the Pcdh. First, we found that CTCF binds in a highly specific manner to the promoter region of several representative genes from the Pcdhα and Pcdhβ families in vitro. Second, knocking down CTCF down-regulated the expression of endogenous active Pcdh genes as well as Pcdh promoter-driven reporter gene. These findings suggest that CTCF binds to the promoters of the clustered Pcdh genes and is acting as a master positive regulator of all the Pcdh genes. Our findings are consistent with previous reports that show that binding of CTCF to the target site is methylation-independent (40,41).

CTCF is a diverse protein and has a unique structure of 11 zinc-finger-DNA binding domains which are conserved among vertebrates. This distinct structure gives CTCF an exceptional degree of flexibility for DNA binding site recognition, which led to the description of CTCF as a ‘multivalent’ transcription factor (42). CTCF is known...
to perform numerous functions including enhancer blocking, X-chromosome inactivation, gene imprinting, monoallelic gene expression and promoter activation or repression (42–46). It has been demonstrated that CTCF can mediate contact between CTCF binding sites, to stabilize intra- and interchromosomal long range interactions to affect transcription (38,47,48). Conditional CTCF Knock out models have demonstrated that CTCF affects transcription in enhancers of both beta and alpha Globin loci (49,50), the HOX cluster (51), the XITE of the X-chromosome (52) and the EBV (53). Another recent study, which investigates the transcriptional activity in CTCF mutant limbs, has shown that the CTCF sites in RNA splicing may regulate the production of specific alternative transcript variants (51).

CTCF is known to interact with diverse protein partners that determine its specific function; hence, the regulation of CTCF activity might be achieved by neighboring factors bound to DNA. These partners factors include the RNA polymerases I, II and III, another zinc finger factor VEZF1 and the factors YY1, SMAD, TR and Oct4. Each of these seems to influence, modulate or determine the function of CTCF (54). Of particular interest to this study is YY1 which was shown to interact with CTCF and to function together in X-chromosome inactivation (52). YY1 is a ubiquitous four-zinc-finger transcription factor that has been implicated in biological processes such as embryogenesis, differentiation, cell proliferation and tumorgenesis (55). Surprisingly, in the course of our research YY1 were found to bind specifically to the SSE–CSE, most likely through a YY1 binding site that partially overlap the CSE of $\alpha 6$, $\alpha 3$ and $\alpha 12$. Although YY1 and CTCF appeared in our experiments as distinct complexes on EMSA, it is possible that under physiological conditions they bind cooperatively to the promoters of Pcdh genes.

Considering that the Pcdh gene cluster contains several promoter elements in tandem, with only one or two are active, and the fact that Pcdh expression is monoallelic, it is possible that in addition to its ability to activate Pcdh transcription, it also has a central role in insulating nearby promoters in the monoallelic expression of Pcdh genes. Depending on the promoter context and cell background, CTCF may repress or activate transcription; however, its repression function predominates. We found that CTCF acts primarily as a positive transcription factor for the Pcdh genes. We cannot conclude from our results that the SSE dictates cell type specific expression. However, if we take into account the fact that CTCF is associated with most of the Pcdh promoters in vivo (56,57) (according to the ChiP-seq data found in the UCSC genome browser), as well as in conserved enhancers HS5-1 (28) and HS16–20 (58), it may well be possible that CTCF also serves as a repressor that acts to silence the inactive Pcdh genes. However in these studies the exact site to which CTCF binds and the functional significance of CTCF for Pcdh expression were not addressed. This repression can be achieved through different conformation of DNA/CTCF-complexes (which CTCF binding to itself) that allow it to form chromatin hubs by selective intra- and interchromosomal interactions bridging together specific subsets of genomic CTCF sites. The divergence of the CTCF-binding sequences (as a result of unique SSEs) can serve as ‘CTCF CODE’ (59,60) encrypting interactions with co-partner in a site specific manner (through the SSEs) and establishing structure-functional 3D organization involved in regulating the expression of individual Pcdh genes. The nature and composition of this co-partner is yet to be determined. Profiling the gene expression pattern of several different cell lines following profiling the DNA binding location of the CTCF and other candidate proteins will give us a predictive mechanistic model for the specific regulation of the Pcdh gene expression.

In summary, the expression of distinct Pcdh mRNAs in individual neurons is regulated by the activation or repression of subsets of promoters preceding individual genes. The choice of a gene included in a Pcdh mRNA appears to be a direct consequence of promoter selection. The mechanism of differential Pcdh promoter activation or repression in individual neurons is therefore likely to be a key step in regulating the cell-specific expression of distinct Pcdh genes. Our work is, therefore, another step forward in the effort to understand the mechanism for Pcdh gene expression, which may have significant implications on mammalian brain development. These findings may also have implications to human neuronal diseases linked to aberrant Pcdh function such as autism (61), bipolar disorder, schizophrenia (62), auditory deficiencies (63) and tumors (64).

SUPPLEMENTARY DATA
Supplementary Data are available at NAR Online: Supplementary Figures 1–5 and Supplementary Data 1–3.

ACKNOWLEDGEMENTS
The authors would like to thank V. Lobanenkov (National Institute of Allergy and Infectious Diseases, National Institutes of Health, Rockville, MD, USA) for generously providing plasmids pET-7.1 and pET-11ZF. We would like to thank Eran Segal for help in bioinformatics analysis and discussions. We would like to thank Ido Amit for the valuable advice and insightful comments, which were very helpful in improving our manuscript. E.S. is the incumbent of The Harry Weinreb Professorial Chair of Computer Science and Biology. R.D. is the incumbent of the Ruth and Leonard Simon Chair of Cancer Research.

FUNDING
Kahn Family Research Center for System Biology of the Human Cell; John von Neumann Minerva Center for the Development of Reactive Systems; European Union (FP7-ERC-AdG). Funding for open access charge: European Union (FP7-ERC-AdG).

Conflict of interest statement. None declared.
REFERENCES

1. Lomvardas,S., Barnea,G., Pisapia,D., Mendelsohn,M., Kirkland,J. and Axel,R. (2006) Intercellular interactions and olfactory receptor choice. *Cell*, 126, 403–413.
2. Schmucker,D., Clemens,J., Shu,H., Worby,C., Xiao,J., Muda,M., Dixon,J. and Zipursky,S. (2000) Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell*, 101, 671–684.
3. Wojtowicz,W.M., Flanagan,J.J., Millard,S.S., Zipursky,S.L. and Clemens,J.C. (2004) Alternative splicing of Drosophila Dscam generates axon guidance receptors that exhibit isoform-specific homophilic binding. *Cell*, 118, 619–633.
4. Hattori,D., Demir,E., Kim,H., Viragh,E., Zipursky,S. and Dickson,B. (2007) Dscam diversity is essential for neuronal wiring and self-recognition. *Nature*, 449, 223–227.
5. Hamada,S. and Yagi,T. (2001) The cadherin-related neuronal receptor family: a novel diversified cadherin family at the synapse. *Neurosci. Res.*, 41, 207–215.
6. Shapiro,L. and Colman,D.R. (1999) The diversity of cadherins and implications for a synaptic adhesion code in the CNS. *Neuron*, 23, 427–430.
7. Serafini,T. (1999) Finding a partner in a crowd: neuronal diversity and synaptogenesis. *Cell*, 98, 133–136.
8. Yagi,T. and Takeichi,M. (2000) Cadherin superfamily genes: functions, genomic organization, and neurologic diversity. *Genes Dev.*, 14, 1169–1180.
9. Enomoto,M. and Jontes,J. (2008) Inhibition of protocadherin-alpha function results in neuronal death in the developing zebrafish. *Dev. Biol.*, 312, 175–187.
10. Phillips,G.R., Tanaka,H., Frank,M., Elste,A., Fidler,L., Benson,D.L. and Colman,D.R. (2003) Gamma-protocadherins are targeted to subsets of synapses and intracellular organelles in neurons. *J. Neurosci.*, 23, 5096–5104.
11. Weiner,J., Wang,X., Tapia,J. and Sanes,J. (2005) Gamma protocadherins are required for synaptic development in the spinal cord. *Proc. Natl Acad. Sci. USA*, 102, 8–14.
12. Kohmura,N., Senzaki,K., Hamada,S., Kai,N., Yasuda,R., Watanabe,M., Ishii,H., Yasuda,M., Mishina,M. and Yagi,T. (1998) Diversity revealed by a novel family of cadherins expressed in neurons at a synaptic complex. *Neuron*, 20, 1137–1151.
13. Wu,Q. and Maniatis,T. (1999) A striking organization of a protocadherin-alpha gene cluster. *Proc. Natl Acad. Sci. USA*, 103, 7971–7976.
14. Wang,X., Weiner,J.A., Levi,S., Craig,A.M., Bradley,A. and Sanes,J.R. (2002) Gamma protocadherins are required for survival of spinal interneurons. *Neuron*, 20, 834–845.
15. Blank,M., Triana-Baltzer,G., Richards,C. and Berg,D. (2004) Alpha-protocadherins are presynaptic and axonal in nociceptive pathways. *Mol. Cell. Neurosci.*, 23, 520–529.
16. Frank,M., Ebert,M., Shan,W., Phillips,G.R., Arndt,K., Colman,D.R. and Kemler,R. (2005) Differential expression of individual gamma-protocadherins during mouse brain development. *Mol. Cell. Neurosci.*, 29, 603–616.
17. Zou,C., Huang,W., Ying,G. and Wu,Q. (2007) Sequence analysis and expression mapping of the rat clustered protocadherin gene repertoires. *Neuroscience*, 144, 579–603.
18. Fukuda,E., Hamada,S., Hasegawa,S., Katori,S., Sanbo,M., Miyakawa,T., Yamamoto,T., Yamamoto,H., Hirabayashi,T. and Yagi,T. (2008) Down-regulation of protocadherin-alpha A isoforms in mice changes contextual fear conditioning and spatial working memory. *Eur. J. Neurosci.*, 28, 1362–1376.
19. Katori,S., Hamada,S., Noguchi,Y., Fukuda,E., Yamamoto,T., Yamamoto,H., Hasegawa,S. and Yagi,T. (2009) Protocadherin-alpha family is required for serotonergic projections to appropriately innervate target brain areas. *J. Neurosci.*, 29, 9137–9147.
20. Sugino,H., Hamada,S., Yasuda,R., Tuji,A., Matsuda,Y., Fujita,M. and Yagi,T. (2000) Genomic organization of the family of CNR cadherin genes in mice and humans. *Genomics*, 63, 73–87.
21. Wu,Q., Zhang,T., Cheng,J., Kim,Y., Grimwood,J., Schmutz,J., Dickson,M., Noonan,J., Zhang,M., Myers,R. et al. (2001) Comparative DNA sequence analysis of mouse and human protocadherin gene clusters. *Genome Res.*, 11, 389–404.
22. Esumi,S., Kakazu,N., Taguchi,Y., Hirayama,T., Sasaki,A., Hirabayashi,T., Koido,T., Kitsukawa,T., Hamada,S. and Yagi,T. (2005) Monoallelic yet combinatorial expression of variable exons of the protocadherin-alpha gene cluster in single neurons. *Nat. Genet.*, 37, 171–176.
23. Kaneko,R., Kato,H., Kutsuneya,M., Esumi,S., Hirayama,T., Hirabayashi,T. and Yagi,T. (2006) Allelic gene regulation of Pcdh-alpha and Pcdh-gamma clusters involving both monoallelic and biallelic expression in single Purkinje cells. *J. Biol. Chem.*, 281, 30551–30560.
24. Zhang,T., Haws,P. and Wu,Q. (2004) Multiple variable first exons: a mechanism for cell- and tissue-specific gene regulation. *Genome Res.*, 14, 79–89.
25. Wu,Q. and Maniatis,T. (2000) Large exons encoding multiple ectodomains are a characteristic feature of protocadherin genes. *Proc. Natl Acad. Sci. USA*, 97, 3124–3129.
26. Tasic,B., Nabinolz,C., Baldwin,K.K., Kim,Y., Rueckert,E.H., Ribich,S.A., Cramer,P., Wu,Q., Axel,R. and Maniatis,T. (2002) Promoter choice determines splice site selection in protocadherin alpha and gamma pre-mRNA splicing. *Mol. Cell.*, 10, 21–33.
27. Wang,X., Su,H. and Bradley,A. (2002) Molecular mechanisms governing Pcdh-gamma gene expression: evidence for a multiple promoter and cis-alternative splicing model. *Genes Dev.*, 16, 1900–1905.
28. Ribich,S., Tasic,B. and Maniatis,T. (2006) Identification of long-range regulatory elements in the protocadherin-alpha gene cluster. *Proc. Natl Acad. Sci. USA*, 103, 19719–19724.
29. Kawaguchi,M., Toyama,T., Kaneko,R., Hirayama,T., Kawamura,Y. and Yagi,T. (2008) Relationship between DNA methylation states and transcription of individual isoforms encoded by the protocadherin-alpha gene cluster. *J. Biol. Chem.*, 283, 12064–12075.
30. Murata,Y., Hamada,S., Morishita,H., Mutoh,T. and Yagi,T. (2004) Interaction with protocadherin-gamma regulates the cell surface expression of protocadherin-alpha. *J. Biol. Chem.*, 279, 49508–49516.
31. Amir-Zilberstein,L., Airnbin,E., Toube,L., Yamaguchi,Y., Handa,H. and Dikstein,R. (2007) Differential regulation of NF-kappaB by elongation factors is determined by core promoter type. *Mol. Cell. Biol.*, 27, 5246–5259.
32. Amir-Zilberstein,L. and Dikstein,R. (2008) Interplay between E-box and NF-kappaB regulation of A20 gene by DRB sensitivity-inducing factor (DSIF). *J. Biol. Chem.*, 283, 1317–1323.
33. Kurakutsi,S., Tiwari,K., Tavoosidana,G., Pugacheva,E., Murrell,A., Zhao,Z., Lobanekov,V., Reik,W. and Ohlsson,R. (2006) CTCF binding at the H19 imprinting control region mediates maternally inherited higher-order chromatin conformation to restrict enhancer access to Igf2. *Proc. Natl Acad. Sci. USA*, 103, 10684–10689.
34. Hancock,A.L., Brown,K.W., Moorwood,K., Moon,H., Holmgren,C., Mardikar,S.H., Dallosso,A.R., Kleinova,E., Loukinov,D. and Ohlsson,R. et al. (2007) A CTCF-binding silencer regulates the imprinted genes AWTI and WT1-AS and exhibits sequential epigenetic defects during Wilms' tumourigenesis. *Hum. Mol. Genet.*, 16, 343–354.
35. Noonan,J.P., Li,J., Nguyen,L., Caoile,C., Dickson,M., Grimwood,J., Schmutz,J., Feldman,M.W. and Myers,R.M. (2003) Extensive linkage disequilibrium, a common 16.7-kilobase deletion, and evidence of balancing selection in the human protocadherin alpha cluster. *Am. J. Hum. Genet.*, 72, 621–635.
36. Noonan,J.P., Grimwood,J., Schmutz,J., Dickson,M. and Myers,R.M. (2004) Gene conversion and the evolution of the protocadherin gene cluster diversity. *Genome Res.*, 14, 354–366.
37. Sukamoto,S., Kabe,Y., Hatakeyama,M., Yamaguchi,Y. and Handa,H. (2009) Development and application of high-performance affinity beads: toward chemical biology and drug discovery. *Chem. Rec.*, 9, 66–85.
38. Handoko,L., Xu,H., Li,G., Nguen,C.Y., Chew,E., Schnapp,M., Lee,C.W., Ye,C., Ping,J.L., Mulawadi,F. et al. (2011) CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat. Genet.*, 43, 630–638.
