Published in final edited form as:

Gene. 2009 January 15; 429(1-2): 74–79. doi:10.1016/j.gene.2008.09.025.

Embryonic nervous system genes predominate in searches for dinucleotide simple sequence repeats flanked by conserved sequences

Donald E. Riley^{a,b,*} and John N. Krieger^{a,b}

^a Department, Urology, University of Washington, Seattle, WA 98195

^b Department, Research, VA Puget Sound Health Care System, Room 517, Building 1, 1660 South Columbian Way, Seattle, WA 98012

Abstract

To study evolution of dinucleotide simple sequence repeats (diSSRs) we searched recently available mammalian genomes for UTR-localized diSSRs with conserved upstream flanking sequences (CFS). There were 252 reported *Homo sapiens* genes containing the repeats (AC)_n, (GT)_n, (AG)_n or (CT)_n in their UTRs including 22 (8.7%) with diSSR-upstream flanking sequences conserved comparing divergent mammalian lineages represented by *Homo sapiens* and the marsupial, *Monodelphis domestica*. Of these 22 genes, 19 had known functions including 18 (95%) that proved critical for mammalian nervous systems (Fishers exact test, $P < 0.0001$). The remaining gene, *Cd2ap*, proved critical for development of kidney podocytes, cells that have multiple similarities to neurons. Gene functions included voltage and chloride channels, synapse-associated proteins, neurotransmitter receptors, axon and dendrite pathfinders, a NeuroD potentiator and other neuronal activities. Repeat length polymorphism was confirmed for 68% of CFS diSSRs even though these repeats were nestled among highly conserved sequences. This finding supports a hypothesis that SSR polymorphism has functional implications. A parallel study was performed on the self-complementary diSSRs (AT)_n and (GC)_n. When flanked by conserved sequences, the self-complementary diSSR (AT)_n was also associated with genes expressed in the developing nervous system. Our findings implicate functional roles for diSSRs in nervous system development.

Keywords

Highly conserved elements; HCEs; Simple sequence repeats; SSRs; dinucleotide simple sequence repeats; diSSRs; microsatellites; nervous system; development

1. Introduction

Simple sequence repeats (SSRs), also termed short tandem repeats or microsatellites, represent some of the most recognizable noncoding sequences. Often, SSRs modulate their associated genes to provide a means for genetic variation with minimal genetic load (Fondon and Garner,

Address for correspondence: Donald Riley, Puget Sound Health Care System (VA) Room 517, Building 1; 1660 South Columbian Way, Seattle, WA 98108 email, dri@u.washington.edu Tel. 206-277-3608; Fax: (206) 764-2164.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

2004; Li et al., 2004; Kashi and King, 2006). Previously, we reported association of the dinucleotide SSRs (diSSRs) (AC)_n and (GT)_n with the 3' untranslated regions (UTRs) of genes encoding membrane functions and transcription factors (Riley and Krieger, 2004). (Sequences are written as they appear relative to the sense-strand of nearby coding sequence.) During database searches, we noticed that mRNAs encoding proteins of similar function (e.g., members of the aquaporin family) sometimes had 3' UTRs with different diSSRs such as (AC)_n or (AG)_n. For this reason, we hypothesized that different diSSR sequences might serve similar functions. This hypothesis led to the prediction that diSSRs might replace one another during evolution of a given UTR.

Comparing orthologous UTRs in different species, dinucleotide SSRs (diSSRs) such as (AC)_n, (GT)_n, (AG)_n and (CT)_n, were found to frequently replaced one another at the same position within the UTR supporting the hypothesis that these repeats function in similar structural roles (Riley, 2004; Riley and Krieger, 2005). Folding algorithms predicted that the potential to form single-stranded loops represented a common structural feature because (AC)_n, (GT)_n, (AG)_n and (CT)_n all lack the ability to form canonical base pairs by themselves. We termed these, "weak-folding repeats."

The diSSRs are present in less than 0.15 % of reported UTRs. Thus, the many precisely localized diSSR replacements that have been observed are unlikely to represent random events. While diSSR replacements proved common, we never observed a triplet repeat replacing a diSSR. Triplet repeats and diSSRs also have different genomic distributions, with the latter more common among noncoding sequences (Toth et al., 2000; Wren et al., 2000; Cordeiro et al., 2001). These observations support separate study of diSSRs since they might represent a distinct functional category.

Compared to the weak-folding diSSRs, the diSSRs (AT)_n and (GC)_n have a somewhat different usage. The self-complementary diSSR (AT)_n ≥ 14 has single-stranded folding potential and was often replaced during evolution by palindromic sequences with similar folding potential (Riley et al., 2007). The diSSR (GC)_n ≥ 14 also has folding potential but is rarely used. There were no human UTRs reported with (GC)_n ≥ 14. Based on its replacement by other folding sequences (AT)_n may function differently from the weak-folding diSSRs.

An excellent past study surveyed all SSRs including di-, tri-, and tetranucleotide repeats at all genomic locations (Toth et al., 2000). In contrast, we restricted study to highly selected SSR groups to allow for the possibility that heterogeneous functions might be obscured if the study's scope were too broad. The current investigation substantially extends this approach by examining UTR-localized diSSRs that have highly conserved flanking sequences. Highly conserved noncoding sequence elements have been termed HCEs (Siepel et al., 2005). We hypothesized that examination of UTR-localized diSSRs, flanked by HCEs, using recently available genome sequences as a resource, might provide new insights into the long-term evolution and function of UTR-localized diSSRs.

Like previous studies of SSRs, past studies of HCEs were broad, encompassing introns, intragenic spacers, gene deserts and UTRs (Siepel et al., 2005). 3' UTRs accounted for an unexpectedly high proportion of HCE bases (11-fold enrichment). Presence of noncoding HCEs predicted genes implicated in development, differentiation and malignancies (Bejerano et al., 2004; Sandelin et al., 2004; Derti et al., 2006). However, there were no clear usage patterns or consistent functional contexts. One possible explanation is that previous broad studies of HCEs may involve heterogeneous functions obscuring usage patterns that might be present in more restricted groups of HCEs. For this reason, we obtained new data using a novel perspective provided by limiting our initial screen to diSSRs flanked by HCEs within mammalian UTRs.

2. Materials and methods

2.1. Database searches and the conserved upstream flanking sequence (CFS) strategy

The CFS strategy was designed to identify a set of diSSRs flanked by HCEs, then study evolution of these sites using recently available genome sequences from a variety of mammalian species. To obtain an initial set of CFS UTRs (those that have diSSRs flanked by HCEs) we used the weak-folding diSSR sequences (AC)_n, (GT)_n, (AG)_n or (CT)_n with $n \geq 14$ to search non-redundant 5' and 3' UTR databases (version, 2.2.1, 4-13-01, 368,154 sequences (<http://www.ba.itb.cnr.it/BIG/Blast/BlastUTR.html>)) for all *Homo sapiens* UTRs that contained such diSSRs. The non-redundant UTR database has been previously described (Pesole et al., 2002). For these searches, we used the Basic Local Alignment Search Tool (BLAST) with low-complexity filtering turned off and default parameters for Expect and Matrix. Using the same parameters, a parallel study of (AT)_n, a “strong folding” diSSR, was done separately because we anticipated (AT)_n might have different usage compared with weak-folding diSSRs. For reasons outlined below, trinucleotide and tetranucleotide repeats did not play major roles in the current study. For the initial search sequences, we accepted no base substitutions within the core repeat ($n \geq 14$) for two reasons: 1. We knew there were hundreds of perfect, core repeats in the UTR database. 2. We were concerned that accepting imperfections might inadvertently capture repeats that were evolving into non-repetitive sequences.

Upstream flanking sequences, consisting of 150 to 350 bp, from each of the 252 UTRs identified were then used to search the opossum genome using BLASTn at the same website. We assumed that upstream flanking sequences conserved comparing a primate and a marsupial would likely recover sequences conserved across a broad range of mammals. Presence of a diSSR in a human UTR, and conservation of the upstream flanking sequence in opossum, represented the only selection criteria applied during CFS strategy searches.

2.2. Coding potential score

The diSSR site sequences were annotated as noncoding sequences in Genbank. However, to allow for possible alternative, unreported transcripts, we used a variation of CSTMiner (Castrignano et al., 2004) to evaluate all diSSR-flanking sequences, both upstream and downstream, for coding potential. The initial CSTMiner screen identifies sequences conserved comparing two species, such as human and mouse, to identify sequences likely to function. The algorithm then assigns coding potential scores (CPS) based on quantification of synonymous and non-synonymous substitutions at the nucleotide level and conservative changes vs. non-conservative ones at the protein level. We followed the classification scheme of Castrignano et al. who categorized sequences as either noncoding sequences ($CPS < 6.74$; probability of coding $< 1\%$), possible coding sequences ($6.74 \leq CPS \leq 7.71$) or coding sequences ($CPS > 7.71$; probability of coding $> 99\%$). For each UTR-localized diSSR conserved in human and opossum, at least 200 b of upstream and downstream SR flanking sequence were pasted into the search window at <http://pentagramma.caspar.it/GenoMinerNew/>. Then, the sequences were compared with the mouse genome and CPS scores obtained. Conservation comparing human and mouse, while less stringent than our CFS strategy, is considered evidence of function. Resulting CPS values are reported in Tables 1 and 2.

2.3. Statistical analyses

The control set was selected from UTR-localized HCE's (Siepel et al., 2005) exhibiting the same range of human-opossum similarity as exhibited by the CFS strategy UTRs. The CFS criterion for minimal similarity was established with that purpose in mind. To evaluate whether the CFS strategy preferentially selected nervous system related genes, we calculated that a

control group of 20 randomly selected genes with known functions provided 80% power to detect a 40% difference with alpha (significance) = 0.01 (two-tailed). To assure that we would select at least 20 randomly selected genes with known functions, we chose to evaluate expression data and functional literature of 100 randomly selected sequences. We configured the Genome Browser at <http://genome.ucsc.edu> (Kent et al., 2002) to identify DNA sequences that intersected with the annotations 'UTR' and 'most conserved.' Of 8,211 UTRs identified, 100 were selected using Minitab® 15.1.0.0, statistical software (State College, PA) configured to produce random data from a uniform distribution of 8,211 integers. Larger numbers of UTRs were not examined because determination of function involved exhaustive literature searches for each gene.

2.4. Determination of function

All genes associated with diSSR-containing UTRs, and the control UTR set, underwent thorough literature and database studies to evaluate existing function and expression data. In addition, phenotypes deduced from targeted null mutations and tissue expression data were reviewed at <http://www.informatics.jax.org/>. Such data were available for the majority of genes identified in the current study. These procedures were not conducive to automation. Tables present brief statements that summarize available data on gene expression and function.

2.5. Polymorphism detection using sequence databases

To test for polymorphism, we searched Genbank, the UTR database and the genome databases referenced above for duplicate entries of the same UTR sequence in different individuals within the species *Mus musculus* and *H. sapiens*. Duplicate gene entries from different individuals were detected by differences in the submitting authors or by differences in the numbers of reported diSSR repeats. This approach facilitated confirmation of polymorphism if different repeat numbers were found but could not exclude polymorphism when duplicate entries had the same number of repeats.

3. Results

3.1. Genes recovered by the CFS strategy

Using the non-redundant UTR database and the Ensemble databases, there were 252 total (AC)n, (GT)n, (AG)n or (CT)n -containing UTRs (where $n \geq 14$) reported for *H. sapiens*. Among these 252 genes, 22 (8.7%) had diSSR-upstream flanking sequences that met the CFS strategy's minimum threshold of $\geq 90\%$ conservation (described in Materials and Methods). Of these 22 genes, 19 had known functions, including 18 (95%) that proved critical in nervous system development and function (Table 1). Except for the kidney developmental gene, *Cd2ap*, the genes identified are strongly expressed in neurons or synaptic vesicles. While both 3' and 5' UTRs were present among the initial 252 UTRs, all 22 CSF strategy UTRs were 3' UTRs.

Among the three genes of unknown function, one, KIAA2026, encodes a protein listed as part of a protein-interacting network involved in inherited ataxia and disorders of Purkinje cell degeneration consistent with a nervous system role (Lim et al., 2006). A second gene with unknown function, AK048412.1, represented a cDNA isolated from a 16-day mouse embryonic head cDNA library but otherwise was uncharacterized. The remaining gene with unknown function lacked data on either function or expression.

Proteins encoded by CFS strategy genes included: axonal and dendritic membrane proteins involved in Ca^{2+} sensing, axon guidance proteins, various neuronal transcription factors, protein kinases, a kinesin motor protein, proteins critical to synapse formation, synaptic vesicle-related activities and *Cd2ap*, essential for podocyte extension during embryogenesis. All of these genes are expressed in the mouse embryo and most have mouse knockout strain

data confirming deleterious effects on central or peripheral nervous system development (Eppig et al., 2005). Some of these genes also function in the adult nervous system and in non-neuronal tissues.

The assumption that diSSR upstream flanking sequences conserved between human and opossum would also be conserved across a broad range of mammals proved largely correct (see companion manuscript).

3.2. (AT)_n and (GC)_n

Previously, we studied (AT)_n-site evolution focusing on folding potential but did not report on function evaluated after the CFS strategy was applied to those sites (Riley et al., 2007). Due to a distinct replacement pattern, we studied (AT)_n separately from the weak-folding diSSRs using the same criteria. In new data generated for the current study, among 230 (AT)_n-containing human UTRs, 14 (6%) had upstream flanking sequences conserved comparing human and opossum (Table 2). Because functional information was limited for CFS strategy (AT)_n genes, Table 2 emphasizes expression sites and effects observed in gene knockout mice. Of the 14 genes, 12 (86%) are expressed in embryonic nervous systems. Two genes, *Bnc1* and *Col4A5*, are expressed in the developing kidney and the latter is also expressed in embryonic brain. Only one CFS strategy (AT)_n gene, *Ahnak*, has no known association with the nervous system or kidney. The diSSR (GC)_n > 14 was absent among reported human UTRs. We conclude that, like the weak-folding diSSRs, (AT)_n and its flanking sequences may have a role in nervous system and kidney development.

3.3. Coding sequence potential

Among the CFS strategy diSSR sites, only the upstream diSSR flanking sequence for *Ches1* had a high probability of encoding protein (Table 1). Upstream (us) or downstream (ds) flanking sequences for *Glp1r* (ds), *Zfx1a* (us), and *Ublcp1* (ds) scored as possible, but unconfirmed coding sequences (CPS = 6.74–7.71). All other diSSR flanking sequences had low CPS scores (< 6.74) with an estimated coding probability of < 1% (Castrignano et al., 2004). We conclude that, in most cases, the high degree of sequence conservation in these regions is most likely explained by functions other than encoding protein.

Among 14 (AT)_n containing UTRs identified by the CFS strategy, there were no diSSR flanking sequences with CPS scores > 7.71 although 6 (43%) of these sequences scored as possible coding sequences (Table 2). Database searches revealed no transcripts. We conclude that the majority of diSSR-flanking sequences identified by the CFS strategy are unlikely to encode proteins.

3.4. Statistical analyses

The number of control UTRs selected was limited by labor intensive literature searches and was based on a statistical power calculation. Absence of diSSRs among the control UTR dataset likely resulted from the fact that diSSRs with ≥ 14 repeats occur in less than 0.15 % of all reported UTRs (Riley and Krieger, 2004).

Among the 100 control UTRs, 80 (80%) had known functions based on individual literature searches and the mouse genome informatics website, www.informatics.jax.org (Eppig et al., 2005). By comparison, among the original 22 CFS UTRs, 19 (85%) had known functions (two-sided $p = 0.7635$ using Fisher's exact test, relative risk (RR) = 1.080; 95% confidence interval (CI): 0.8902 to 1.309). Thus, availability of functional information proved similar among conserved UTRs with and without diSSRs.

Then, we compared the functional details for UTRs with and without diSSRs. The 80 control UTRs whose genes had known functions included 30 (38%) whose functions were related to nervous system development or function. In contrast, the 19 UTRs whose genes had known functions identified using the CFS strategy included 18 (95%) related to nervous system development or function (two-sided $P < 0.0001$ using Fisher's exact test, $RR = 2.526$, $CI: 1.867$ to 3.418). There was less functional information available for (AT) $_n$ containing UTRs with conserved diSSR flanking sequences. However, the association with nervous system expression proved significant ($P < 0.001$). Finally, when we searched a non-redundant UTR database using title-line terms, neuronal|neuron|axon|dendrite|brain| only 272 (0.01%) human UTRs were recovered out of 29,416 total human UTRs (Chi square $P < 0.0001$).

3.5. Detection of polymorphism

Strong conservation of diSSR flanking sequences raised a question whether diSSRs in such sequence environments were polymorphic or perhaps rigidly conserved within a species. Of the 22 CFS strategy UTRs, the *Vsn1l* diSSR flanking sequences were the most conserved with strong similarity both upstream and downstream of the repeat region in vertebrates including an amphibian and three fish species examined.

To investigate possible polymorphisms, we searched Genbank for *Vsn1l* sequence data from alternative individuals (Altschul et al., 1997). The *M. musculus Vsn1l* cDNA reported by Shiri et al. under accession number D21165 has (AC) $_{21}$. In contrast, the *M. musculus Vsn1l* cDNA reported by Strausberg et al. under accession number BC046226 has (AC) $_{16}$. Two additional independent *M. musculus Vsn1l* cDNAs plus the genome sequence at Ensemble had either (AC) $_{16}$ or (AC) $_{21}$. Among independent human *Vsn1l* encoding sequences (accessions numbers, NW_001838767.2 and NM_003385.4) (AC) $_{17}$ and (AC) $_{22}$ were found respectively. These observations confirmed that among mice and humans, the (AC) $_n$ repeat is polymorphic. Whether this diSSR is as polymorphic as (AC) $_n$ in less conserved sequence environments will require further study.

Using the same approach, we searched for polymorphisms in the other CFS strategy diSSRs. Of 22 CFS strategy diSSRs 15 (68%) proved to be polymorphic. The 7 remaining diSSRs (*Glp1r*, *Pou4f2*, *Ches1*, *Rreb1*, *Nlk*, *Ublep1* and *Faslg*) lacked sufficient database entries to resolve the question of polymorphism.

4. Discussion

4.1. Hypothesis to explain selection of embryonic nervous system genes by the CFS strategy

Although we anticipated that the CFS strategy might identify a distinct group of genes, the sequences recovered raised two questions: "Why embryonic genes?" and, "Why nervous system genes?" In retrospect, some possible, albeit unproven answers became obvious. At the molecular level, CFS-based identification of embryonic genes echoes Baer's 1828 observation, that, "All developing vertebrates appear very similar shortly after gastrulation. It is only later in development that the special features of class, order, and finally species emerge" (Baer, 1828). At the molecular level, embryos of different species contain both homologous and non-homologous molecules (Gilbert, 2000).

We hypothesize that selection of embryonic genes by the CFS strategy may reflect, in part, a pool of UTRs that serve embryonic functions fixed early in mammalian evolution and then conserved. Expression of molecules that are very similar among species is likely to be more common early in development than at later stages when most species differences emerge. The HCE-diSSR sites studied here represent a small fraction of HCEs likely to be of future interest in this regard (Siepel et al., 2005).

Selection of nervous system genes by the CFS strategy is reminiscent of the fact that most triplet SSR expansion diseases are neurological disorders. While genomic distributions of diSSRs and triplet SSRs are generally different, there is some overlap. At least six triplet repeat expansion diseases involve triplet repeats in 5' or 3' UTRs. Genes *DM1*, *SCA8* and *HDL2* associated with myotonic dystrophy, spinal cerebellar ataxia and Huntingtons disease-like 2 respectively involve 3' UTR repeat expansions (Nithianantharajah and Hannan, 2007; Orr and Zoghbi, 2007). Currently, there is no evidence for diSSR expansion disease. However, diSSR lengths are kept within a relatively narrow range suggesting selective forces operating on repeat lengths (Li et al., 2004). For genes and UTRs critical in development, diSSR expansion may lead to embryonic death perhaps contributing to absence of clinical disease in neonates and later in life.

Evidence that SSRs modulate developmental genes, particularly brain developmental genes, has been reviewed (Nithianantharajah and Hannan, 2007). In contrast to triplet repeats, which have been intensively investigated for the last fifteen years, diSSRs remain relatively obscure even though their potential significance for gene regulation was recognized much earlier (Hamada et al., 1984; Comings, 1998; Orr and Zoghbi, 2007).

An intronic diSSR (AC)_n was shown to modulate gene expression with shorter alleles leading to increased *Egfr* gene expression. In breast cancer, (AC)_n region amplifications were present in patients who had allelic imbalances (Tidow et al., 2003). There is evidence that an SSR in the 5' UTR of the prairie vole *Avpr1a* gene generates diversity in brain and sociobehavioral traits (Hammock and Young, 2005). Length variation in the diSSR (GA)_n led to changes in gene expression, brain distribution and behavior. In humans this site is occupied by a diSSR couple (two diSSRs abutted together).

The one non-neuronal gene identified in the original CFS strategy gene collection, *Cd2ap*, is critical for kidney podocyte development (Eppig et al., 2005). Like neurons, podocytes are highly polar cells with elaborate cytoplasmic projections (for a scanning electron micrograph, please see supplementary material). Podocytes also produce functional synaptic vesicles and express activities previously thought specific to neurons (Rastaldi et al., 2006).

4.2. Levels of selection for nervous system genes

Statistical analyses demonstrated that the initial screen for UTRs with diSSRs exerted a substantial preference for nervous system genes. The data suggest that selection based on conservation alone (i.e. selection for the control set of HCEs) also exerted a preference for nervous system genes. Given these findings, it may not be surprising that combining the two constraints led to the strong selection of nervous system genes in the current study. Potential functions of diSSRs and flanking sequences will be examined in more detail in the accompanying paper.

4.3. Alternative approaches

Our search strategy was asymmetric. We chose to start by comparing a late branching lineage with an early branching lineage to potentially minimize the number of false leads. By starting with long, perfect repeats in humans, we found these sites often to be less perfect in other species (please see companion paper). In an alternative approach, one might start with longer perfect diSSRs in each genome and then search the other genomes. In theory, this more difficult, iterative approach would lead to a greater number of highly conserved diSSR sites. Even more information might be gained by examining narrower ranges of mammals or other selected organisms. The diSSR flanking sequences are usually unique in their genomes enabling one to locate these sites when present in other genomes. In theory, tetranucleotide repeats could be studied by a similar approach. Among human UTRs that had tetranucleotide repeats, we found

none with conserved flanking sequences in opossum. A parallel study of tetranucleotide repeats will likely require a narrower range of species.

5. Conclusions

Without intentionally selecting for function, the CFS strategy identified genes predominantly involved in mammalian nervous system function and development. These studies resulted in a focused group of genes supporting a functional role for some diSSRs with conserved flanking sequences in the embryologic development and function of highly polar cells such as neurons and kidney podocytes. These novel findings also suggest a potential link between the functions of certain highly conserved noncoding sequences (HCEs) and some of the most variable noncoding sequences (diSSRs).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Abbreviations

CFS	conserved flanking sequence
HCE	highly conserved element
UTR	untranslated region
diSSRs	dinucleotide simple sequence repeats
us	upstream
ds	downstream

References

- Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;25:3389–3402. [PubMed: 9254694]
- Baer, KEv. *Entwicklungsgeschichte der Thierte: Beobachtung und Reflexion*. Borntrager; Konigsberg: 1828.
- Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D. Ultraconserved elements in the human genome. *Science* 2004;304:1321–1325. [PubMed: 15131266]
- Castillo PE, Janz R, Sudhof TC, Tzounopoulos T, Malenka RC, Nicoll RA. Rab3A is essential for mossy fibre long-term potentiation in the hippocampus. *Nature* 1997;388:590–593. [PubMed: 9252190]
- Castrignano T, Canali A, Grillo G, Liuni S, Mignone F, Pesole G. CSTminer: a web tool for the identification of coding and noncoding conserved sequence tags through cross-species genome comparison. *Nucleic Acids Res* 2004;32:W624–627. [PubMed: 15215464]
- Chang BH, Mukherji S, Soderling TR. Calcium/calmodulin-dependent protein kinase II inhibitor protein: localization of isoforms in rat brain. *Neuroscience* 2001;102:767–777. [PubMed: 11182241]

- Chen B, Schaevitz LR, McConnell SK. Fezl regulates the differentiation and axon targeting of layer 5 subcortical projection neurons in cerebral cortex. *Proc Natl Acad Sci U S A* 2005;102:17184–17189. [PubMed: 16284245]
- Chubykin AA, Liu X, Comoletti D, Tsigelny I, Taylor P, Sudhof TC. Dissection of synapse induction by neuroligins: effect of a neuroligin mutation associated with autism. *J Biol Chem* 2005;280:22365–22374. [PubMed: 15797875]
- Comings DE. Polygenic inheritance and micro/minisatellites. *Mol Psychiatry* 1998;3:21–31. [PubMed: 9491809]
- Cordeiro GM, Casu R, McIntyre CL, Manners JM, Henry RJ. Microsatellite markers from sugarcane (*Saccharum s*) ESTs cross transferable to *erianthus* and *sorghum*. *Plant Sci* 2001;160:1115–1123. [PubMed: 11337068]
- Derti A, Roth FP, Church GM, Wu CT. Mammalian ultraconserved elements are strongly depleted among segmental duplications and copy number variants. *Nat Genet* 2006;38:1216–1220. [PubMed: 16998490]
- Doudney K, Murdoch JN, Paternotte C, Bentley L, Gregory S, Copp AJ, Stanier P. Comparative physical and transcript maps of approximately 1 Mb around loop-tail, a gene for severe neural tube defects on distal mouse chromosome 1 and human chromosome 1q22–q23. *Genomics* 2001;72:180–192. [PubMed: 11401431]
- During MJ, et al. Glucagon-like peptide-1 receptor is involved in learning and neuroprotection. *Nat Med* 2003;9:1173–1179. [PubMed: 12925848]
- Eppig JT, et al. The Mouse Genome Database (MGD): from genes to mice--a community resource for mouse biology. *Nucleic Acids Res* 2005;33:D471–475. [PubMed: 15608240]
- Falls DL. Dasm1: a receptor that shapes neuronal dendrites and turns on silent synapses? *Sci STKE* 2005;2005:pe10. [PubMed: 15755926]
- Fondon JW 3rd, Garner HR. Molecular origins of rapid and continuous morphological evolution. *Proc Natl Acad Sci U S A* 2004;101:18058–18063. [PubMed: 15596718]
- Gilbert, SF. *Developmental Biology*. Sunderland, MA: Sinauer Assoc. Inc.; 2000.
- Graf ER, Zhang X, Jin SX, Linhoff MW, Craig AM. Neurexins induce differentiation of GABA and glutamate postsynaptic specializations via neuroligins. *Cell* 2004;119:1013–1026. [PubMed: 15620359]
- Hamada H, Seidman M, Howard BH, Gorman CM. Enhanced gene expression by the poly(dT-dG).poly(dC-dA) sequence. *Mol Cell Biol* 1984;4:2622–30. [PubMed: 6098815]
- Hammock EA, Young LJ. Microsatellite instability generates diversity in brain and sociobehavioral traits. *Science* 2005;308:1630–1534. [PubMed: 15947188]
- Huber TB, et al. Bigenic mouse models of focal segmental glomerulosclerosis involving pairwise interaction of CD2AP, Fyn, and synaptopodin. *J Clin Invest* 2006;116:1337–1345. [PubMed: 16628251]
- Kashi Y, King DG. Simple sequence repeats as advantageous mutators in evolution. *Trends Genet* 2006;22:253–259. [PubMed: 16567018]
- Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. *Genome Res* 2002;12:996–1006. [PubMed: 12045153]
- Landau AM, et al. Defective Fas expression exacerbates neurotoxicity in a model of Parkinson's disease. *J Exp Med* 2005;202:575–581. [PubMed: 16129703]
- Laterza OF, Modur VR, Crimmins DL, Olander JV, Landt Y, Lee JM, Ladenson JH. Identification of Novel Brain Biomarkers. *Clin Chem* 2006;1713–1721. [PubMed: 16858073]
- Li YC, Korol AB, Fahima T, Nevo E. Microsatellites within genes: structure, function, and evolution. *Mol Biol Evol* 2004;21:991–1007. [PubMed: 14963101]
- Lim J, Hao T, Shaw C, Patel AJ, Szabo G, Rual JF, Fisk CJ, Li N, Smolyar A, Hill DE, Barabasi AL, Vidal M, Zoghbi HY. A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. *Cell* 2006;125:801–814. [PubMed: 16713569]
- Lin J, et al. Defects in adaptive energy metabolism with CNS-linked hyperactivity in PGC-1alpha null mice. *Cell* 2004;119:121–135. [PubMed: 15454086]

- Martin LJ, Chen K, Liu Z. Adult motor neuron apoptosis is mediated by nitric oxide and Fas death receptor linked by DNA damage and p53 activation. *J Neurosci* 2005;25:6449–6459. [PubMed: 16000635]
- Mathisen PM, Johnson JM, Kawczak JA, Tuohy VK. Visinin-like protein (VILIP) is a neuron-specific calcium-dependent double-stranded RNA-binding protein. *J Biol Chem* 1999;274:31571–31576. [PubMed: 10531361]
- Miki H, Setou M, Kaneshiro K, Hirokawa N. All kinesin superfamily protein, KIF, genes in mouse and human. *Proc Natl Acad Sci U S A* 2001;98:7004–7011. [PubMed: 11416179]
- Mok H, Shin H, Kim S, Lee JR, Yoon J, Kim E. Association of the kinesin superfamily motor protein KIF1B α with postsynaptic density-95 (PSD-95), synapse-associated protein-97, and synaptic scaffolding molecule PSD-95/discs large/zona occludens-1 proteins. *J Neurosci* 2002;22:5253–5358. [PubMed: 12097473]
- Nithianantharajah J, Hannan AJ. Dynamic mutations as digital genetic modulators of brain development, function and dysfunction. *Bioessays* 2007;29:525–35. [PubMed: 17508392]
- Orr HT, Zoghbi HY. Trinucleotide repeat disorders. *Annu Rev Neurosci* 2007;30:575–621. [PubMed: 17417937]
- Pesole G, Liuni S, Grillo G, Licciulli F, Mignone F, Gissi C, Saccone C. UTRdb and UTRsite: specialized databases of sequences and functional elements of 5' and 3' untranslated regions of eukaryotic mRNAs. Update 2002. *Nucleic Acids Res* 2002;30:335–340. [PubMed: 11752330]
- Puschel AW, Betz H. Neurexins are differentially expressed in the embryonic nervous system of mice. *J Neurosci* 1995;15:2849–2856. [PubMed: 7722633]
- Rastaldi MP, et al. Glomerular podocytes contain neuron-like functional synaptic vesicles. *Faseb J* 2006;20:976–978. [PubMed: 16585060]
- Ray SK, Nishitani J, Petry MW, Fessing MY, Leiter AB. Novel transcriptional potentiation of BETA2/NeuroD on the secretin gene promoter by the DNA-binding protein Finb/RREB-1. *Mol Cell Biol* 2003;23:259–271. [PubMed: 12482979]
- Riley DE. Simple repeat replacements support similar functions of distinct repeats in inter-species mRNA homologs. *Gene* 2004;328C:17–24. [PubMed: 15019980]
- Riley DE, Jeon JS, Krieger JN. Simple repeat evolution includes dramatic primary sequence changes that conserve folding potential. *Biochem Biophys Res Commun* 2007;355:619–625. [PubMed: 17320822]
- Riley DE, Krieger JN. Short tandem repeats are associated with diverse mRNAs encoding membrane-targeted proteins. *Bioessays* 2004;26:434–44. [PubMed: 15057941]
- Riley DE, Krieger JN. Short tandem repeat (STR) replacements in UTRs and introns suggest an important role for certain STRs in gene expression and disease. *Gene* 2005;344:203–211. [PubMed: 15656986]
- Samady L, Faulkes DJ, Budhram-Mahadeo V, Ndisang D, Potter E, Brabant G, Latchman DS. The Brn-3b POU family transcription factor represses plakoglobin gene expression in human breast cancer cells. *Int J Cancer* 2006;118:869–878. [PubMed: 16152597]
- Sandelin A, Bailey P, Bruce S, Engstrom PG, Klos JM, Wasserman WW, Ericson J, Lenhard B. Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. *BMC Genomics* 2004;5:99. [PubMed: 15613238]
- Shi SH, Cox DN, Wang D, Jan LY, Jan YN. Control of dendrite arborization by an Ig family member, dendrite arborization and synapse maturation 1 (Dasm1). *Proc Natl Acad Sci U S A* 2004;101:13341–13345. [PubMed: 15340157]
- Shirataki H, Kaibuchi K, Sakoda T, Kishida S, Yamaguchi T, Wada K, Miyazaki M, Takai Y. Rabphilin-3A, a putative target protein for smg p25A/rab3A p25 small GTP-binding protein related to synaptotagmin. *Mol Cell Biol* 1993;13:2061–2068. [PubMed: 8384302]
- Siepel A, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 2005;15:1034–1050. [PubMed: 16024819]
- Sons MS, Plomp JJ. Rab3A deletion selectively reduces spontaneous neurotransmitter release at the mouse neuromuscular synapse. *Brain Res* 2006;1089:126–134. [PubMed: 16631140]
- Stobrawa SM, et al. Disruption of CIC-3, a chloride channel expressed on synaptic vesicles, leads to a loss of the hippocampus. *Neuron* 2001;29:185–196. [PubMed: 11182090]

- Su JH, Anderson AJ, Cribbs DH, Tu C, Tong L, Kesslack P, Cotman CW. Fas and Fas ligand are associated with neuritic degeneration in the AD brain and participate in beta-amyloid-induced neuronal death. *Neurobiol Dis* 2003;12:182–93. [PubMed: 12742739]
- Thorpe CJ, Moon RT. nemo-like kinase is an essential co-activator of Wnt signaling during early zebrafish development. *Development* 2004;131:2899–2909. [PubMed: 15151990]
- Tidow N, Boecker A, Schmidt H, Agelopoulos K, Boecker W, Buerger H, Brandt B. Distinct amplification of an untranslated regulatory sequence in the egfr gene contributes to early steps in breast cancer development. *Cancer Res* 2003;63:1172–1178. [PubMed: 12649171]
- Toth G, Gaspari Z, Jurka J. Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res* 2000;10:967–981. [PubMed: 10899146]
- Triboli C, Robledo RF, Lufkin T. The murine fork head gene Foxn2 is expressed in craniofacial, limb, CNS and somitic tissues during embryogenesis. *Mech Dev* 2002;118:161–163. [PubMed: 12351180]
- Wren JD, et al. Repeat polymorphisms within gene regions: phenotypic and evolutionary implications. *Am J Hum Genet* 2000;67:345–356. [PubMed: 10889045]
- Yoshikawa M, et al. CLC-3 deficiency leads to phenotypes similar to human neuronal ceroid lipofuscinosis. *Genes Cells* 2002;7:597–605. [PubMed: 12059962]
- Zhao C, et al. Charcot-Marie-Tooth disease type 2A caused by mutation in a microtubule motor KIF1Bbeta. *Cell* 2001;105:587–597. [PubMed: 11389829]

Table 1

Genes recovered with UTR-localized weak-folding diSSRs and CFS criteria

Category	Gene	Function	CPS
Neuronal membrane proteins	<i>Vsn1l</i>	calcium sensing; CNS injury biomarker (Mathisen et al., 1999; Laterza et al., 2006)	< 6.74
	<i>Sema6D</i>	axon targeting in cerebral cortex (Chen et al., 2005)	< 6.74
	<i>Glp1r</i>	learning and neuroprotection (During et al., 2003)	ds, 6.74–7.71
	<i>Pou4f2</i>	axon guidance (Samady et al., 2006)	< 6.74
	<i>Zfhlx1a</i>	essential for axon development (Eppig et al., 2005)	us, 6.74–7.71
Neuronal transcription factors	<i>Ches1</i>	strong expression CNS development (Tribioli et al., 2002)	us > 7.71
	<i>Ppargc1a</i>	CNS hyperactivity; receptor coactivator; (Lin et al., 2004)	< 6.74
	<i>Rreb1</i>	potentiator of neurogenic factor NeuroD (Ray et al., 2003)	< 6.74
	<i>Kcnip</i>	transcriptional repressor involved in pain modulation (Eppig et al., 2005)	< 6.74
	<i>Nlk</i>	midbrain patterning (Thorpe and Moon, 2004)	< 6.74
Neuronal kinases or phosphatases	<i>Camk2n1</i>	frontal cortex, hippocampus, colliculus (Chang et al., 2001)	< 6.74
	<i>Ublcp1</i>	expressed in embryonic brain and other (Eppig et al., 2005)	< 6.74
Neuronal microtubule motor	<i>Kif1b</i>	brain size/synaptic vesicle development (Miki et al., 2001; Zhao et al., 2001; Mok et al., 2002)	< 6.74
Neuronal apoptosis	<i>Faslg</i>	adult motor neuron degeneration/apoptosis (Su et al., 2003; Landau et al., 2005; Martin et al., 2005)	< 6.74
Synapse formation and function	<i>Nrxn1 (imp)</i>	neurotransmitter release and synapse formation (Puschel and Betz, 1995; Graf et al., 2004; Chubykin et al., 2005)	< 6.74
	<i>Cln3 (imp)</i>	chloride channel; loss of hippocampus; motor chord deficit (Stobrawa et al., 2001; Yoshikawa et al., 2002)	< 6.74
	<i>Rab3A</i>	neuromuscular synapse and learning (Shirataki et al., 1993; Castillo et al., 1997; Sons and Plomp, 2006)	< 6.74
Dendritic membrane protein	<i>Igsf9</i>	dendrite development/arborization/learning (Doudney et al., 2001; Shi et al., 2004; Falls, 2005)	< 6.74
Kidney podocyte development	<i>Cd2ap</i>	essential for podocyte extension/development (Eppig et al., 2005; Huber et al., 2006)	< 6.74

Table 2

Genes recovered with UTR-localized (AT)n and CFS criteria

Gene	Expression sites	Function (homozygous null mutant phenotype)	CPS
<i>Adcyap1</i>	embryonic brain	behavioral anomalies	us, 6.74–7.71
<i>Ahnak</i>	keratinocyte plasma membrane protein	no known abnormalities	us, 6.74–7.71
<i>Arhgap21</i>	embryonic brain and postnatal kidney	Unknown	< 6.74
<i>Atp2b2</i>	embryonic brain and heart	slow growth and deafness, associated with cerebellar abnormalities	< 6.74
<i>Bach2</i>	cerebral cortex	impaired b cell differentiation and reduced b cell numbers	< 6.74
<i>Bnc1</i>	mesonephros, urogenital ridge and other tissues	failure to divide past two cell stage	< 6.74
<i>Cacnb2</i>	embryonic neural retina and heart	vision and nervous system abnormalities	< 6.74
<i>Col4A5</i>	embryonic brain and kidney glomerular tuft	premature death, proteinuria, elevated blood urea nitrogen, and kidney glomerular and tubular malformations (Alport Syndrome in humans)	us, 6.74–7.71
<i>Ddx6</i>	embryonic brain and trunk	Unknown	< 6.74
<i>Rab11fip3</i>	embryonic and postnatal neural retina	Unknown	< 6.74
<i>Rib140</i>	embryonic brain and other tissues	Unknown	< 6.74
<i>Senp1</i>	embryonic brain and other tissues	widespread cell death	us, 6.74–7.71
<i>Zfp608</i>	embryonic brain and other tissues	Unknown	us, 6.74–7.71
<i>Znf516</i>	embryonic brain and other tissues	unknown	us, 6.74–7.71