

Distribution and Patterns of CNSs in a *Caenorhabditis* gene family
Mark Bieda and Colleen T. Webb

Abstract

Conserved non-coding sequences (CNSs) are conserved sequences in genomes and often contain *cis*-regulatory elements such as transcription factor binding sites. The KCNK gene family proteins represent an ancient system of ionic regulation found in nearly all metazoan cell types. A conservative set of 33 genes from this family was derived by using structural/biophysical criteria. Here, we investigate the distribution and prevalence of CNSs derived from alignments of the genomes of *Caenorhabditis elegans* and *Caenorhabditis briggsae* in this family, which includes >300 kb of non-coding DNA (ncDNA) in >350 regions (intronic and intergenic). Overall, there were ~14 CNSs/gene and ~10% of ncDNA nucleotides were in CNSs. A significant portion of CNSs (~30%) were found in introns and most genes possessed at least one intronic CNS. Generally, CNSs were spatially clumped, had higher GC content than ncDNA outside of CNSs, and displayed significant transition bias. For both introns and intergenic regions, CNSs were found at similar normalized prevalence, as measured by percent coverage of ncDNA and number of CNSs per kb of ncDNA, and CNS prevalence was correlated with region length. Intronic CNSs were preferentially distributed towards the 5' end, and to a lesser extent, the 3' end of the gene, with significantly fewer in middle introns. These results suggest that intronic CNSs comprise a major population of CNSs from *C. elegans*–*C. briggsae* and that the processes governing intronic and intergenic CNSs may be similar.

Introduction

One basic division of genomic DNA is into coding and non-coding sequence. For metazoans generally, non-coding DNA (ncDNA) occupies the majority of sequence space (*C. elegans*: 75.7%, *C. briggsae* 76.7%, *Homo sapiens* 98.3%; (Stein et al. 2003; Taft and Mattick 2003)). Interestingly, it has recently been claimed that the genomic percentage of ncDNA correlates well with organismal complexity (Taft and Mattick 2003). Although repeat elements comprise a significant portion of ncDNA for most metazoans, ncDNA harbors a large variety of functional elements, including regulatory elements such as insulators, enhancers, silencers, promoters, and matrix attachment regions, and structural genomic elements such as signals for chromosome assembly and DNA replication. It is obviously important to catalog the types, prevalence, patterns, and roles of these elements for understanding the genome. However, with the exception of repeats, the computational difficulty in predicting these elements in ncDNA has made study of them difficult, and hence there is little understanding of their large-scale patterns and prevalence.

One way to gain insight into ncDNA is to examine conservation of segments of ncDNA between species. These segments are often termed “phylogenetic footprints” or “conserved non-coding sequences (CNSs)”. CNSs potentially give information on many aspects of ncDNA. CNSs represent evolutionarily conserved segments of ncDNA, so information on their properties should supply insight into general aspects of global, large-scale genomic processes of gene regulation and molecular evolution in ncDNA. Because large scale patterns in coding DNA (e.g. positioning of genes in chromosomes; Roy et al. 2002) appear to be an important organizing principle in genomes, large scale patterns in ncDNA are also expected to be important. In mammalian genomes, a recent striking result is that a subset of CNSs are hyperconserved over long stretches of evolutionary time, probably signifying particularly important roles for these elements (Bejerano et al. 2004; Dermitzakis et al. 2003). Also, overall patterns/prevalence of CNSs may give insight into large scale differences in regulation between organism groups. For example, there has been recent commentary that grasses appear to have many fewer CNSs, as measured by various parameters, than mammals, which has been interpreted to mean that grasses may have a simpler regulatory system (Inada et al. 2003). Furthermore, CNSs have been effectively used as guides to functionally-relevant sites in ncDNA, including in *C. elegans* (Salkoff et al. 2001). In addition, CNSs derived from multiple species alignments have proven valuable in ascertaining transcription factor binding sites (Cliften et al. 2003; Kellis et al. 2003) and there are extensive efforts in this direction currently (e.g. Thomas et al. 2003).

KCNK potassium channels are two-pore region potassium channels currently under intensive investigation. This protein family represents an ancient, ubiquitous, system of ionic regulation in metazoans; members of this gene family are differentially expressed in a large number of metazoan cell types. In *C. elegans*, there are ~40 putative members of this gene family (“twk” channels; (Kunkel et al. 2000; Salkoff et al. 2001)), which are known to be selectively targeted and these channels may play a critical role in “tuning” neurons (Salkoff et al. 2001). More generally, KCNK channels are thought to be important in establishing cellular membrane potential and conductance and, for neurons, are known targets of neuromodulators (Goldstein et al. 2001). In addition, in mammals, these channels are important clinically in that they are targets of some anesthetics (Goldstein et al. 2001) and transcriptional dysregulation of these channels plays a role in some cancers (Pei et al. 2003). Hence, changes in KCNK channels, whether in protein properties or regulation, may have large impacts on cell physiology, and this ancient, ubiquitous gene family provides an interesting example case for examining evolutionary changes in CNSs and regulation in the context of a range of physiological processes.

Model organisms have been shown to be valuable in gaining insight into evolutionarily conserved cellular processes. Therefore, it seems reasonable to expect that a study of KCNK genes in these organisms could give insight into general principles of regulation in this ancient family. *C. elegans* offers many advantages for study of this gene family, including a very high quality genome database, extensive general information, a large variety of well-established

experimental techniques, relatively fast generation time and fast experimental time, and an increasing ability to do high-throughput work. In addition, the divergence time of *C. elegans* and *C. briggsae* is long enough for mutational saturation to occur (Shabalina and Kondrashov 1999; Kent and Zahler 2000), making these species ideal for investigation of selectively conserved functional sequences.

In this work, we examined CNSs in the KCNK channel family in *C. elegans* by using comparative sequence analysis of *C. elegans* and *C. briggsae*. In addition to the physiological rationale for examining transcriptional regulation of these channels, the moderate size of this gene family, the known structural and functional information, and the differential distribution offer advantages for addressing more general genomics questions. Our work had several overlapping goals, including (i) establishing a set of CNSs in this gene family for further experimental and computational studies, including (a) studies attempting to link specific CNSs to specific spatiotemporal expression patterns and (b) studies analyzing changes in CNSs with gene duplication and divergence; (ii) examining CNS prevalence and distribution patterns in *Caenorhabditis*, in particular with reference to previous work on patterns in intergenic regions in *C. elegans* (Kent and Zahler 2000; Shabalina and Kondrashov 1999; Webb et al. 2002) and previous work in other organisms (Bergman and Kreitman 2001; Inada et al. 2003; Kaplinsky et al. 2002; Keightley and Gaffney 2003); (iii) examining patterns in introns, for which there is limited information in *C. elegans*; (iv) using CNSs to attempt to extract common regulatory motifs in this gene family (Wasserman and Sandelin 2004). In addition to these primary goals, we also present a partial curation of the set of *C. elegans* KCNK genes based on biophysical/structural criteria.

Results

Determination of Elegans Set of KCNK Channels

KCNK proteins are constituents of two-pore/four-transmembrane potassium channels that have a TM1-P1-TM2-TM3-P2-TM4 structure, where TM1-4 are transmembrane alpha helices and P1 and P2 are pore regions (Goldstein et al. 2001). From structural information on the Kcsa potassium channel (Doyle et al. 1998) and known properties of KCNK channels (Goldstein et al. 2001) we derived three basic criteria to evaluate putative KCNK family members for inclusion in the final set of analyzed channels: (i) existence of two GXG motifs, where X is any amino acid (AA) (there must be one GXG per pore region); (ii) an aromatic amino acid (F/W/Y) 9 residues upstream of the first G in each of these found GXG motifs; (iii) assuming that each transmembrane alpha helix is at least 20 AAs in length and given the TM1-P1-TM2-TM3-P2-TM4 structure, the following should hold: the first GXG must be at least 20 AAs from the N-terminus to accommodate TM1; there must be >40 AAs between the two GXG motifs to accommodate TM2-TM3; there must be at least 20 AAs after the second GXG motif to accommodate TM4. Figure 1A shows an example of these features for one gene in the set (K01D12.4). We used this information to filter a putative list of *C. elegans* KCNK channels (from Salkoff et al. 2001) to form a conservative set of KCNK family sequences in *elegans*. We found that 33 of 41 putative proteins met these criteria (Table 1).

Basic Properties of OWEN Alignments

OWEN is a relatively new sequence alignment tool specifically designed to find collinear local alignments of sequence using a hierarchical alignment approach; i.e. sequences that with high similarity are aligned first – then sequences with progressively lower similarity are aligned using remaining sequence (Ogurtsov et al. 2002; Roytberg et al. 2002).

Figure 1B shows a representation of an OWEN alignment of K01D12.4 in *C. elegans* and *C. briggsae*. We extracted the sequence including the KCNK gene, the 5' and 3' intergenic regions, and the exons of the non-KCNK genes flanking these regions ("Exons" track). In the "OWEN exons" track, we display portions of alignment segments that were within the boundaries of exons (other pieces are not shown for clarity). Note that there is excellent coverage of the K01D12.4 exons, good coverage of the flanking exon on the left (K01D12.5), and poor coverage of the flanking exon on the right (K01D12.15). Data from wormbase (<http://www.wormbase.org>) for this region of the *elegans* genome indicates that, in agreement with our alignments, K01D12.4 and K01D12.5 both align to the same *C. briggsae* contig (cb25.fpc4470), while the poor conservation of K01D12.15 is expected because it aligns best to a different *C. briggsae* contig (cb25.fpc3752).

As a larger test of our procedures, we examined the percentage of nucleotides in exons of *C. elegans* ("exonic nucleotides") that were included in our aligned sequences. We found that our alignments overlapped 94 +/- 6 % (median 96%; range 77-100%) of exonic nucleotides on a per gene basis. In addition, 242 of the 326 total exons (74%) in the 33 gene set had 100% of exonic nucleotides in alignments and 95.7% of exons had >60% of nucleotides in alignments. Only 6 annotated *C. elegans* exons had zero nucleotides in alignments. Hence, OWEN appeared effective in detecting homologous segments.

General CNS Characteristics

Figure 1B ("OWEN CNSs" track) shows CNSs generated by our alignment procedure after eliminating exons and filtering for statistical significance (see METHODS). This plot shows several features that we investigated in this study, including the presence of both intronic and intergenic CNSs, a bias in intronic CNSs to introns near gene ends, greater number of intergenic vs intronic CNSs, and a "clumpy" pattern of CNS distribution.

In the set of 33 genes, there was ~308 kb of non-coding sequence. Overall, 10.5% of ncDNA was covered in alignments (known ncRNAs were excluded from this, but represented an extremely small contribution). Two genes (T01B4.1 and W06D12.5) did not have any CNSs, but these genes together only accounted for ~12 kb of sequence, so the percentage coverage is little affected by their inclusion. We found a total of 476 CNSs, for an average of ~14/gene and ~1.6

CNSs per kb of ncDNA. These 476 were distributed so that 134 were in the 293 total introns in the set of genes, for an average of 0.46 CNSs/intron, and 342 were in the 66 intergenic regions, for an average of ~5 CNSs/intergenic region. CNS length in *elegans* and *briggsae* was almost identical (Fig 2A; *elegans*: 68 +/- 39 nts (range 21-340); *briggsae*: 69 +/- 39 nts (range 21-340)).

Similarity of sequence in an alignment is defined as the number of matches of nts in that sequence divided by the total length of the sequence. CNS similarity for *elegans* and *briggsae* sequences, on average, was identical (Fig 2B; 82 +/- 7 %, range 65%-100% for *elegans* and *briggsae*). This measure of similarity includes both random matches and matches by selective constraint. To calculate the actual fraction of nucleotides in a footprint that were selectively constrained, we used the methods of Shabalina and Kondrashov (1999) (see METHODS). Again, the average values and distribution of selective constraint was essentially identical for *elegans* and *briggsae* sequences in CNSs (Fig 2C; *elegans*: 75% +/- 11% (range 48-100%); *briggsae*: 75% +/- 10% (range 48-100%)). Therefore, as expected, our CNSs represent islands of well-aligned sequence in which most nucleotides are selectively constrained.

The ratio of transitions to transversions (referred to here as “mutation ratio”) has been found to generally be greater than one for exons and much less than one for intergenic and intronic regions taken as a whole (e.g. Webb et al. (2002) found it to be 0.46 for intergenic sequence outside of CNSs). 459 of 476 CNSs had defined mutation ratios (i.e. >0 transversions). We investigated the mutation ratio in CNSs and found it to be 1.1 +/- 0.95 (range 0-8). However, inspection of the distribution revealed a “fat tail” in that 30 of the 459 footprints had quite high ratios (>=3). Elimination of these led to an average of 0.92 +/- 0.62. In both sets of data (the full set and the set with >=3 eliminated), the median was about 0.80 (0.83 for full set; 0.80 for set with large values eliminated). The measured distribution (from full set of 459) is significantly different from the null expectation of 0.5 ($p < E-34$), and indicates a significant overabundance of transitions, consistent with an island of conserved sequence under selection.

Clumping of CNSs

To examine potential clumping of CNSs, we measured the interval between the end of one CNS and the start of the next for CNSs in the same region (“interhit interval”; same intron or intergenic region; Fig 3A). Distances were only computed between CNSs; distances to neighboring exons were not included in this distribution. Because CNSs are generally small in relation to regions, if positioning of CNSs is random, this should conform to a Poisson process (see Webb et al. 2002). A histogram of the interhit interval demonstrates that there are more than expected at short distances, and less than expected at moderate distances, indicating clumping, which also appears evident in the example shown in Fig 1B.

We also examined potential inhomogeneous coverage in regions close to the KCNK 5' and 3' ends (Fig 3B). In agreement with Webb et al. (2002), we found that there was an upward shift in the range 1-50 nts into the 3' end. However, while Webb et al. (2002) observed a similar rise on the 5' end, this was not very pronounced in our data. In addition, unlike Webb et al. (2002) who found a relatively stable plateau of coverage from nts 60-300 on the 5' end, we observed potential signs of additional structure (i.e. “peaks” and “valleys” in coverage) in CNS coverage in the 5' region proximal to the gene. In part, these may be due to the relatively small number of intergenic regions in our sample (only those with intergenic regions >300 nts included: n=28 for 3' intergenic; n=33 for 5' intergenic), but may also be explained by the relative specificity of the KCNK gene family in comparison to the random sample of Webb et al. (2002).

GC Content

Functional regions are often GC-rich compared to non-functional regions. For *C. elegans* (*C. elegans* “hybrid set” from Stein et al. (2003)), exonic GC content is 42.8% and non-exonic GC content is ~32%. For each ncDNA segment (i.e. intergenic segment or intron) that possessed CNSs, we counted the number of G + C nts in all CNSs and the number of G + C nts in all non-CNS portions of the segment. Three regions were excluded from the analysis; 2 had no non-CNS nucleotides and one had a single non-CNS nucleotide. From these numbers of G + C nts, we computed a single GC percentage for the CNS portion of the segment and a single GC percentage for the non-CNS portion of the segment, and we computed the difference in these percentages: (CNS portion mean: 36.4 +/- 8.3%, median: 38.3, range 8.5-58; non-CNS portion

mean: 31.7 +/- 3.8% median: 31.8, range 20.4 - 45.6; $p < E-6$ paired t-test). The distributions of values for CNS and non-CNS portions are shown in Figure 4A; the distribution of differences (i.e. for each segment, the difference = (%GC of CNS portion) – (%GC of non-CNS portion)) is shown in Figure 4B. On a per-region basis, the GC content in CNSs was usually higher than in non-CNSs (69/89 regions; 78%). Hence, CNSs are sequences with enhanced GC content.

Intergenic Region CNSs

We analyzed 66 intergenic regions (2 x 33 genes) comprising ~216 kb of sequence. 18 of 66 (27%) intergenic regions had zero CNSs; only two genes (T01B4.1 and W06D12.5) had both intergenic regions devoid of CNSs, and these genes lacked CNSs entirely. 342 of 476 CNSs were in intergenic regions for an average of 5.2 +/- 7.3 CNSs/intergenic region (range 0-42).

As a general feature of intergenic regions, in accord with a previous study in *C. elegans* (Webb et al. 2002), we found that intergenic region length was correlated with the number of 5' gene ends flanking the region (Fig 5A). The number of CNSs was also correlated with the number of 5' ends (Fig 5B). As expected from these two relations, the number of footprints was also correlated with the region length (data not shown).

Overall, 11% of intergenic ncDNA was in CNSs and there was a global rate of ~1.6 CNSs per kb of intergenic DNA. On average per intergenic region, 12 +/- 14 % (range 0 to 73) of intergenic region nucleotides were included in CNSs. However, percent coverage did not correlate with the number of 5' ends ($r=0.06$; $p>0.6$) and did not correlate with region length ($r=-.05$; $p>0.6$).

A previous study (Webb et al. 2002) found a weak ($r^2=.09$) but statistically significant ($p<0.0001$) correlation between percent of GC in footprints and the number of 5' ends. In this set of KCNK genes, we failed to find this correlation ($r=0.035$; $p>0.5$).

Intronic CNSs

The distribution of CNSs in introns was highly inhomogeneous. Only 20/33 genes possessed an intronic CNS and of these 20 genes, 18 had 1-9 CNSs, while 2 genes had >30 CNSs (Fig 6A). Only 15% of introns had CNSs (44/293), as compared to 73% of intergenic regions. Overall, in the 33 genes, there was a total of ~92.5 kb of intronic DNA (~1.5 CNSs per kb of intronic DNA), of which 8.7 kb are in alignments for a global percentage of 9.4%.

However, like intergenic CNSs, the number of intronic CNSs was positively correlated with the summed length of introns in a gene (Fig 6B; $r=0.85$; $p<E-9$), a slightly higher correlation than that found with intergenic region length ($r=0.61$). In accord with this, the total number of nts covered by CNSs was also strongly correlated with total intron length ($r=0.85$; $p<E-9$).

We investigated the distribution of CNSs among introns. To accommodate the different numbers of introns in different genes (range 5-21), we used a normalization procedure (see METHODS). We found that the distribution of intronic CNSs was biased toward the 5' ends and, to a lesser extent, the 3' ends, forming a U-shaped distribution (Fig 6C).

Motif Analysis

A number of approaches have been developed to search for shared regulatory sites controlling sets of genes. Commonly, the 5' intergenic region of a set of genes is supplied to these algorithms, which essentially look for short sequence words that are statistically overrepresented. Because CNSs have been shown to be good guides to regulatory sequences, instead of using the full 5' intergenic region of the KCNK channels, we used CNSs from the 5' intergenic region.

To test this methodology, we studied an example case – genes in touch receptors thought to be regulated by *mec-3* (Zhang et al. 2002). We generated CNSs for these genes using an identical methodology to that used for KCNK genes, except that we used p-value cutoffs of $p<0.0001$, $p<0.01$, $p<0.05$, $p<0.99$, with the reasoning one can argue that either more relaxed ($p<0.05$, $p<0.99$) or, conversely, more stringent ($p<0.0001$) parameters may yield better definition of important shared motifs. We then used two approaches: one based on Gibbs sampling (Thompson et al. 2003) and one based on exhaustive determination of statistically overrepresented words (Sinha and Tompa 2002, 2000, 2003). For Gibbs sampling, we examined p-values of $p<0.0001$, $p<0.01$, $p<0.05$, $p<0.99$ and three word lengths (6,7,8), for a total of twelve conditions. We found that the core hexamer motif of ATGCAT (actually UNC-86 core binding

motif - see METHODS) was found in 8/12 tests, with no clear pattern of failure. Similarly, using an approach that determines significance based on overrepresented words in reference to a background model of *C. elegans* intergenic sequences (YMF and findexplanators; (Sinha and Tompa 2002)), we tested these conditions and also tested allowing either zero or 2 ambiguous nucleotide characters, leading to a total of 24 conditions tested (4 p-values x 3 word lengths x two values for ambiguous nts). We examined the top 5 motifs found in each condition (top 5 determined by findexplanators). ATGCAT (or, for tests with ambiguous characters, motifs that encompass ATGCAT such as ATGSRT) was found in 16/24 tests; here, there was a clear pattern – the motif was always found with word lengths of 6 and 7 and never found with word lengths of 8. This example generally implies that extraction of motifs may be robust across a reasonably wide range of parameters; with the caveat that longer word lengths may be problematic in some cases.

We searched for common motifs among KCNK genes using the same algorithms and parameter ranges as used for the *mec-3* regulated genes. One observation is that the motif from the *mec-3* regulated genes (ATGCAT and variants with ambiguous nucleotides) was never observed under any of the tested conditions. Gibbs sampling produced only low-complexity motifs. These low-complexity motifs were prevalent in KCNK CNSs but were also prevalent in *mec-3* set CNSs, indicating that they are unlikely to be meaningful. YMF/findexplanators found higher complexity motifs, but these motifs were only shared by small sets of KCNK genes. These data are consistent with the lack of a general regulatory system for KCNK genes.

Discussion

In this study, we examined CNSs in the set of *C. elegans* KCNK genes. We had dual purposes in undertaking this study: from a physiological point of view, to provide more insight into regulation of KCNK genes, and from a genomics perspective, to use this family to investigate general questions surrounding conservation in genomes. The most striking results of our study are from our investigation of intronic CNSs. To our knowledge, there has been little systematic investigation of intronic CNS patterns/prevalence in *C. elegans*. Intronic CNSs formed a significant population of CNSs (~28%). We found some significant differences in CNS patterns in introns vs. intergenic regions. First, while nearly all genes (~94%) had at least one intergenic CNS, many genes lacked intronic CNSs (~61% had intronic CNSs). More strikingly, while most intergenic regions had CNSs (~73%), few introns had CNSs (~15%). Also, intronic CNSs were most often found in introns near the gene ends. However, in other respects, intronic CNSs seemed similar to intergenic CNSs, including a correlation between length and number of CNSs, and the relative coverage of ncDNA by CNSs and per kb prevalence. A second significant contribution of this study was confirmation and extension of previous results regarding patterning of CNSs in intergenic regions (Webb et al. 2002); we found qualitative agreement in patterning, but some significant quantitative differences in prevalence (we found a larger number of CNSs, on average, in intergenic regions, and a lower percentage of overall coverage; see below for details). Given that it has been argued that these prevalence parameters may be used to evaluate overall genomic regulatory complexity (Inada et al. 2003; Kaplinsky et al. 2002), it will be important to resolve these numerical differences. Finally, these results provide a database of CNSs for experimental study to determine regulatory regions and, from a genomics perspective, to provide insight into CNS changes with gene duplication.

KCNK Family Determination

We used three basic criteria for establishing KCNK genes: GXG motif presence in two copies; a properly located upstream aromatic amino acid; sufficient and appropriate residues to form the surrounding 4 transmembrane alpha helices. A general pattern is that the GXG motif is very well conserved in the first pore region and less well conserved in the second pore region (Chapman et al. 2001; Goldstein et al. 2001). Variations in these pore regions may be important for the selectivity of K⁺ over Na⁺ in these channels, and hence may affect physiological function (Chapman et al. 2001). It is clear that there is also significant variation in KCNK channels in open-channel probability and conductance, which also may be important for differential physiological roles (Goldstein et al. 2001). Our insistence on the GXG motif for the second pore region may have artificially eliminated a few valid KCNK channels, and our set represents a “conservative” set. We chose relatively stringent criteria potentially leading to elimination of valid family members with the logic that, for our purposes, loss of a small proportion of the family would probably have little effect on our conclusions, while introduction of non-family members might obscure subtle common patterns.

Determination of CNSs

To discover CNSs, we used OWEN to align sequences followed by filtering of OWEN pieces, with exons removed, using Karlin-Altschul statistics at the $p < 0.01$ level (Karlin and Altschul 1993). One concern is that failure to detect CNSs (low sensitivity) or detection of non-CNSs (low specificity) would distort our results. A recent study concluded that local alignment tools can detect islands of conservation with both high sensitivity and high specificity in the divergence range of *elegans-briggsae* (Pollard et al. 2004); however, OWEN was not examined in that study. As one basic test of the sensitivity of OWEN, we examined the ability of OWEN to detect conserved exons. Generally, there was excellent coverage of exons, a result particularly striking given that there may be some differences in exon lengths between *briggsae* and *elegans*. However, it is important to note that non-coding sequence islands will generally have different properties than coding sequence islands, so this test must be interpreted with caution. It does seem reasonable, based on conclusions of simulations (Pollard et al. 2004) to assume that our

CNS prediction is accurate enough to maintain the validity of our qualitative results, but further studies will be required to assess this.

Comparison to Previous Work in C. elegans

We analyzed our data to investigate previously reported patterns in *elegans-briggsae* CNS in intergenic regions (Webb et al. 2002). Aside from some differences in analyses performed, important differences in data and methods between that study and this one include (1) we analyzed data from a single family of genes (66 intergenic regions), while Webb et al. (2002) chose 142 intergenic regions randomly; (2) we analyzed introns (293 introns); (3) we used OWEN followed by statistical filtering while Webb et al. (2002) used WABA (a global alignment tool) followed by statistical filtering.

Our data is consistent with Webb et al. (2002) in many respects, including average CNS length (Webb et al. (2002): 62 (range 22-292); this study: 68 (range 21-340)); similarity of CNSs (Webb: 80% (range 64 – 100); this study: 82% (range 65 – 100)); selective constraint of CNSs (Webb: 71% (range 46 – 100); this study: 75% (48 – 100)); the existence of relatively strong and similar magnitude correlations between intergenic length and number of footprints (Webb: $r=0.7$; this study $r=0.6$); the existence of relatively weak and similar magnitude correlations between intergenic length and number of 5' gene ends (Webb: $r^2=0.10$; this study: $r^2=0.11$); a significant difference in GC content of CNSs vs non-CNS sections of ncDNA; the existence of some intergenic regions lacking footprints. We failed to find the weak correlation between intergenic region GC content and number of 5' ends as detected in Webb et al. (2002), but the very weak nature of this correlation may make it difficult to detect in a small data set. These basically consistent findings indicate that these qualitative observations from the relatively small samples used in these studies (142 intergenic regions in Webb et al. (2002); 66 in this study) may generalize to the whole genome. Also, these similarities imply that the use of OWEN vs WABA may not have significant qualitative effects on these findings, and also suggest that these results, at least on a qualitative level, may be algorithm-independent.

There were some seemingly significant differences, numerically, in results. We found more CNSs in an average intergenic region than Webb et al. (2002) (Webb: 2.3 (range 0-19); this study: 5.2 (range 0-42)). However, we both found a wide range in number of CNSs/intergenic region. Webb et al. (2002) also found a larger percentage, overall, of intergenic region coverage by CNSs (~20%) than we did (10% for all ncDNA; 11% for intergenic regions). However, as expected from the large variation in CNSs/intergenic region, this percentage varied a great deal. These numerical differences could be due to algorithms used for alignments (WABA vs. OWEN), type of genes in the set, and/or differences in intergenic region length.

Comparison to CNSs in other species

Given that CNSs should reflect patterns of functional conservation in genomes, it is interesting to compare our results to results from other groups of organisms. Inada et al. (2003) suggest that cross-group comparisons of CNS prevalence and patterns can be used as an indicator of genomic regulatory complexity.

Inada et al. (2003) examined CNSs in 52 orthologous genes in maize and rice using BLAST 2 SEQUENCES, and compared their results to mouse-human alignments using similar parameters. Overall, 27% of genes had zero CNSs, and on average there were relatively few CNSs (~4/gene). CNSs were relatively small (most <20 nts), and overall, CNSs covered only 1-4% of ncDNA. To compare to mouse-human CNSs (which appear to have a very similar divergence to maize-rice), Inada et al. (2003) examined a small set of human-mouse orthologues using similar parameters, and in some cases examined the mouse-human orthologues of their maize-rice genes. A consistent finding was the CNSs were significantly larger, more frequent, and denser in mouse-human as compared to maize-rice, which was interpreted to indicate that, at minimum, humans and mice have a more complex regulatory system than maize and rice. One particularly interesting finding was that CNS prevalence seemed to be a function of general gene class. Inada et al. (2003) divided their 52 genes into three classes (enzymes, structural genes, gene regulatory genes) and found that one class (gene regulatory genes) had significantly more CNSs, at a significantly higher overall coverage of ncDNA, than the others. In comparison, our CNS prevalence data for our *briggsae-elegans* comparisons (~14 CNSs/gene, 6% of genes with

zero CNSs, average size of 68 nt, and overall ncDNA coverage of ~10%) are consistently and significantly larger than Inada et al. (2003). Given that KCNK genes are most likely similar to “enzymes” in the Inada classification, these differences are particularly strong, and according to Inada et al. (2003), could be interpreted as indicating that *Caenorhabditis* has a significantly more complex regulatory system than grasses. However, there is a methodology difference between our study and that of (Inada et al. 2003) (OWEN vs. BLAST 2 SEQUENCES), and simulations indicate that, at the approximate divergence range of maize-rice, some algorithms may tend to underreport CNSs (Pollard et al. 2004).

Bergman and Kreitman (2001) investigated CNSs in 40 genes in *Drosophila melanogaster* – *Drosophila virilis* alignments using a filtered dotplotting approach. The 40 genes were chosen on the basis of known experimental evidence for regulatory sequences in their ncDNA. Two important qualitative points of agreement with our study are that intergenic and intronic CNSs were similar and that the transition/transversion ratio is close to 1 for CNSs. Overall, 22% of genes had zero CNSs, with an average of ~31 CNSs/gene. CNSs were small (average of 24 nt) and had a similarity of ~93%. CNSs covered ~22-26% of ncDNA. However, these were ungapped CNSs, which may partially explain, the greater prevalence and smaller size of CNSs in Bergman and Kreitman (2001) vs our study. However, the more restrictive definition of CNSs in Bergman and Kreitman (2001) should lead to less overall coverage of ncDNA sequence space than in our study, but the opposite is observed, indicating that there could be greater conservation of ncDNA, overall, in *Drosophila* than in *Caenorhabditis*.

There have been several studies of mouse-human alignments, which have found that as much as ~23% of intronic nts may be in CNSs (Jareborg et al. 1999). However, using a different approach, Keightley and Gaffney (2003) find that only 9-12% of mouse intronic nucleotides are under selective constraint. Generally, there has not been extensive investigation of CNS patterning in mouse-human alignments, and recent simulations indicate that it may be difficult to determine CNSs with high specificity and high sensitivity at the divergence range of mouse-human (Pollard et al. 2004).

Motifs

We were unable to detect any complex motifs that were present in a large number of KCNK genes. There are several potential explanations for this, including failure of CNSs to contain shared motifs, failure of the algorithms to detect shared motifs, or the lack of existence of a common regulatory system. This third biological possibility is consistent with the divergent spatial regulation of these genes and the large variation in EST prevalence (Salkoff et al. 2001).

Use Of KCNK Gene Family For Gene Duplication Studies

KCNK family genes provide significant advantages for general study of genomic processes of conservation in ncDNA in the context of gene duplication. First, these channels are part of a moderate size family with recent gene duplications (Salkoff et al. 2001). Second, these genes are under intense current physiological study, hence, understanding of differential functional roles is expected (Goldstein et al. 2001). Third, given that they are in *C. elegans*, there can be extensive study of regulation in this model organism. Fourth, there is known heavily differential patterns in expression; for example, one gene (ZK1067.5) is expressed in only sensory neurons, while others (e.g. F19D8.1) are expressed in several cell types (Salkoff et al. 2001). In a limited study, CNSs have already been shown to be guides to ncDNA sequences regulating expression (Salkoff et al. 2001). Fifth, sequencing of several more species of *Caenorhabditis* is underway, providing further ways to study regulation of these channels, at minimum from a comparative sequence analysis viewpoint. Such cross-species analyses have proven useful in understanding evolution of regulation of developmental genes in *Drosophila*. Given all of these advantages, it seems that KCNK family could provide an excellent test case for understanding changes in gene function and regulation with gene duplication.

Future Directions

There are several possible future extensions and expansions of this work. Sequences of additional *Caenorhabditis* species (scheduled for 2004; <http://www.genome.gov/10002154>) will enable determination of multi-species CNSs associated with KCNK genes, providing perhaps

greater insight into evolution of these segments and potentially dividing the set into highly and less-highly conserved pieces. Highly conserved CNSs should be carefully studied. Second, these CNSs provide a database of elements for experimental study as regulatory elements. There has already been a limited but successful use of this approach for this gene family (Salkoff et al. 2001). Third, extension of these analyses to other gene families in *C. elegans* and, more generally, in particular a larger analyses of introns will be important. One issue is whether gene families differ systematically in CNS patterns. There is already some evidence that different general classes of genes differ systematically in CNS prevalence in grasses (Inada et al. 2003). Finally, an analysis of CNS patterns in this gene family with respect to gene relatedness will be very interesting, particularly with respect to whether closely related genes have more similar CNSs (or patterns of CNSs) than more distantly related genes (see above section).

Methods

Source of sequences

C. elegans chromosome sequences were downloaded in final form in 12/2002 and 1/2003 from Genbank (Genbank format files) and wormbase (GFF-format files and FASTA-format chromosomes; ftp://ftp.wormbase.org/pub/wormbase/elegans-current_release/CHROMOSOMES/). Both file types were processed (automatically via Perl programs) to extract a single sequence piece containing the KCNK gene, the surrounding intergenic regions, and the exon of the non-KCNK gene that flanked each intergenic region (see Fig 1B for example). *C. briggsae* genome was from the agp8.cb25 assembly (Stein et al. 2003). Initially, the genomic sequence of each KCNK gene was blasted against the entire set of *C. briggsae* constructs to simply determine which fpc contained the *elegans* gene. Subsequently, information from wormbase (<http://www.wormbase.org>) was used to determine correct fpc constructs for use with OWEN.

OWEN alignments

OWEN (Ogurtsov et al. 2002) was downloaded and run under Linux. Alignments were performed individually via interaction with the program through the graphical interface, beginning with $p < E^{-8}$ and a match requirement of 12 successive matches required to start alignments and masking of repeats. After progressively relaxing these criteria, alignments were finished at the $p < 0.01$ level with 6 successive matches required to start alignments and repeat masking “turned off”. These “raw” OWEN alignments were then filtered as previously described (Webb et al. 2002) by a separate Perl program to yield alignments valid at $p < 0.01$.

Repeats

Gene sequences (which included the flanking exon, intergenic regions, and KCNK genes) were filtered by repeatmasker (<http://www.repeatmasker.org/>). Then, the found repeats were compared (via a Perl program) to our CNSs (which were at $p < 0.01$, as indicated above). It was found that only 14 of 476 CNSs overlapped repeats and in some cases, this overlap was minimal (e.g. 1 nt of overlap in a 61 nt CNS). These 14 CNSs that overlapped repeats were not preferentially distributed toward any gene. Hence, inclusion of these 14 pieces would have minimal effect on our analyses, and we therefore did not eliminate or trim these 14 pieces to eliminate overlaps with repeats.

mec-3 genes

The list of *mec-3* regulated genes was taken from Figure 3 of Zhang et al. (2002). Genes from this list were used if (1) there was a corresponding probable *briggsae* orthologue (2) the positions of the putative binding site, as indicated in Figure 3 of Zhang et al. (2002), could be found using patser (<http://www.ucmb.ulb.ac.be/bioinformatics/rna-tools>), as used by Zhang et al. (2002). This led to a final set of 18 genes. Zhang et al. (2002) quote AATGCAT as the core portion of the motif (actually the core portion of the UNC-86 binding site – see Zhang et al. (2002)), but the initial ‘A’ is poorly conserved (see Fig 3A of Zhang et al. (2002)); hence, to accommodate words of length 6, in our analyses the motif target was considered the core portion ATGCAT.

Calculation of Selective Constraint

We used the approach of Shabalina and Kondrashov (1999) to calculate selective constraint; essentially, this approach adjusts the similarity value for random matches to derive the selective constraint value. For the shorter sequence in an alignment, selective constraint (C_{short}) is:

$$(1) C_{\text{short}} = (s-r)/(1-r)$$

where s is the similarity and r is the probability of a random match.

For the longer sequence in an alignment, selective constraint (C_{long}) is:

$$(2) C_{\text{long}} = (L_{\text{short}}/L_{\text{long}}) * C_{\text{short}}$$

where L_{short} is the length of the shorter sequence and L_{long} is the length of the longer sequence.

These statistics, and all other CNS statistics, were computed in a Perl program.

Normalization of Intron Number

Genes had a wide range of total numbers of introns (5-21). To allow comparison between relative intron numbers, we mapped them onto the interval $(0,1)$ by simply dividing intron number by total introns. For example, if intron 3 possessed a CNS in a gene with 6 introns, we gave it a value of 0.5 (i.e. $3/6$). For our analyses, we simply scored introns as having presence of CNSs or absence; we did not take into account number of CNSs. Taking into account the actual numbers of CNSs/intron produces a distribution with stronger bias toward ends (data not shown).

Author's Contributions

MB and CTW conceived of study and planned overall approach. MB produced sequence extraction from databases, curation of KCNK genes, OWEN alignments, majority of data analysis and interpretation, drafting of manuscript and figures, and all Perl programming. CTW produced OWEN alignments and data analysis and interpretation.

Acknowledgments

MB was a Santa Fe Institute Postdoctoral Fellow during this project; CTW was a Santa Fe Institute Postdoctoral Fellow and also a faculty member at Colorado State University – Fort Collins. We gratefully acknowledge the support of the Santa Fe Institute. CTW is please to acknowledge the support of the David and Lucile Packard Foundation through grant 2000-01690 to the Santa Fe Institute and the National Science Foundation, though grant DEB-0083566 to S.A. Levin. We also acknowledge the work of Christopher Lewis in contributing several software tools and some analysis of motif data. We also acknowledge Richard Aldrich (Stanford University) for advice on curation of KCNK genes.

References

- Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, and Haussler D.** Ultraconserved elements in the human genome. *Science* 304: 1321-1325, 2004.
- Bergman CM and Kreitman M.** Analysis of conserved noncoding DNA in *Drosophila* reveals similar constraints in intergenic and intronic sequences. *Genome Res* 11: 1335-1345, 2001.
- Chapman ML, Krovetz HS, and VanDongen AM.** GYGD pore motifs in neighbouring potassium channel subunits interact to determine ion selectivity. *J Physiol* 530: 21-33, 2001.
- Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, Waterston R, Cohen BA, and Johnston M.** Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* 301: 71-76, 2003.
- Dermitzakis ET, Reymond A, Scamuffa N, Ucla C, Kirkness E, Rossier C, and Antonarakis SE.** Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs). *Science* 302: 1033-1035, 2003.
- Doyle DA, Morais Cabral J, Pfuetzner RA, Kuo A, Gulbis JM, Cohen SL, Chait BT, and MacKinnon R.** The structure of the potassium channel: molecular basis of K⁺ conduction and selectivity. *Science* 280: 69-77, 1998.
- Goldstein SA, Bockenhauer D, O'Kelly I, and Zilberberg N.** Potassium leak channels and the KCNK family of two-P-domain subunits. *Nat Rev Neurosci* 2: 175-184, 2001.
- Inada DC, Bashir A, Lee C, Thomas BC, Ko C, Goff SA, and Freeling M.** Conserved noncoding sequences in the grasses. *Genome Res* 13: 2030-2041, 2003.
- Jareborg N, Birney E, and Durbin R.** Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. *Genome Res* 9: 815-824, 1999.
- Kaplinsky NJ, Braun DM, Penterman J, Goff SA, and Freeling M.** Utility and distribution of conserved noncoding sequences in the grasses. *Proc Natl Acad Sci U S A* 99: 6147-6151, 2002.
- Karlin S and Altschul SF.** Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc Natl Acad Sci U S A* 90: 5873-5877, 1993.
- Keightley PD and Gaffney DJ.** Functional constraints and frequency of deleterious mutations in noncoding DNA of rodents. *Proc Natl Acad Sci U S A* 100: 13402-13406, 2003.
- Kellis M, Patterson N, Endrizzi M, Birren B, and Lander ES.** Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423: 241-254, 2003.
- Kent WJ and Zahler AM.** Conservation, regulation, synteny, and introns in a large-scale *C. briggsae*-*C. elegans* genomic alignment. *Genome Res* 10: 1115-1125, 2000.
- Kunkel MT, Johnstone DB, Thomas JH, and Salkoff L.** Mutants of a temperature-sensitive two-P domain potassium channel. *J Neurosci* 20: 7517-7524, 2000.
- Ogurtsov AY, Roytberg MA, Shabalina SA, and Kondrashov AS.** OWEN: aligning long collinear regions of genomes. *Bioinformatics* 18: 1703-1704, 2002.
- Pei L, Wiser O, Slavin A, Mu D, Powers S, Jan LY, and Hoey T.** Oncogenic potential of TASK3 (*Kcnk9*) depends on K⁺ channel function. *Proc Natl Acad Sci U S A* 100: 7803-7807, 2003.
- Pollard DA, Bergman CM, Stoye J, Celniker SE, and Eisen MB.** Benchmarking tools for the alignment of functional noncoding DNA. *BMC Bioinformatics* 5: 6, 2004.
- Roy PJ, Stuart JM, Lund J, and Kim SK.** Chromosomal clustering of muscle-expressed genes in *Caenorhabditis elegans*. *Nature* 418: 975-979, 2002.
- Roytberg MA, Ogurtsov AY, Shabalina SA, and Kondrashov AS.** A hierarchical approach to aligning collinear regions of genomes. *Bioinformatics* 18: 1673-1680, 2002.
- Salkoff L, Butler A, Fawcett G, Kunkel M, McArdle C, Paz-y-Mino G, Nonet M, Walton N, Wang ZW, Yuan A, and Wei A.** Evolution tunes the excitability of individual neurons. *Neuroscience* 103: 853-859, 2001.
- Shabalina SA and Kondrashov AS.** Pattern of selective constraint in *C. elegans* and *C. briggsae* genomes. *Genet Res* 74: 23-30, 1999.
- Sinha S and Tompa M.** Discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res* 30: 5549-5560, 2002.
- Sinha S and Tompa M.** A statistical method for finding transcription factor binding sites. *Proc Int Conf Intell Syst Mol Biol* 8: 344-354, 2000.

Sinha S and Tompa M. YMF: A program for discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res* 31: 3586-3588, 2003.

Stein LD, Bao Z, Blasiar D, Blumenthal T, Brent MR, Chen N, Chinwalla A, Clarke L, Clee C, Coghlan A, Coulson A, D'Eustachio P, Fitch DH, Fulton LA, Fulton RE, Griffiths-Jones S, Harris TW, Hillier LW, Kamath R, Kuwabara PE, Mardis ER, Marra MA, Miner TL, Minx P, Mullikin JC, Plumb RW, Rogers J, Schein JE, Sohrmann M, Spieth J, Stajich JE, Wei C, Willey D, Wilson RK, Durbin R, and Waterston RH. The Genome Sequence of *Caenorhabditis briggsae*: A Platform for Comparative Genomics. *PLoS Biol* 1: E45, 2003.

Taft R and Mattick J. Increasing biological complexity is positively correlated with the relative genome-wide expansion of non-protein-coding DNA sequences. *Genome Biology* 5: P1, 2003.

Thomas JW, Touchman JW, Blakesley RW, Bouffard GG, Beckstrom-Sternberg SM, Margulies EH, Blanchette M, Siepel AC, Thomas PJ, McDowell JC, Maskeri B, Hansen NF, Schwartz MS, Weber RJ, Kent WJ, Karolchik D, Bruen TC, Bevan R, Cutler DJ, Schwartz S, Elnitski L, Idol JR, Prasad AB, Lee-Lin SQ, Maduro VV, Summers TJ, Portnoy ME, Dietrich NL, Akhter N, Ayele K, Benjamin B, Cariaga K, Brinkley CP, Brooks SY, Granite S, Guan X, Gupta J, Haghghi P, Ho SL, Huang MC, Karlins E, Laric PL, Legaspi R, Lim MJ, Maduro QL, Masiello CA, Mastrian SD, McCloskey JC, Pearson R, Stantripop S, Tiongson EE, Tran JT, Tsurgeon C, Vogt JL, Walker MA, Wetherby KD, Wiggins LS, Young AC, Zhang LH, Osoegawa K, Zhu B, Zhao B, Shu CL, De Jong PJ, Lawrence CE, Smit AF, Chakravarti A, Haussler D, Green P, Miller W, and Green ED. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* 424: 788-793, 2003.

Thompson W, Rouchka EC, and Lawrence CE. Gibbs Recursive Sampler: finding transcription factor binding sites. *Nucleic Acids Res* 31: 3580-3585, 2003.

Wasserman WW and Sandelin A. Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet* 5: 276-287, 2004.

Webb CT, Shabalina SA, Ogurtsov AY, and Kondrashov AS. Analysis of similarity within 142 pairs of orthologous intergenic regions of *Caenorhabditis elegans* and *Caenorhabditis briggsae*. *Nucleic Acids Res* 30: 1233-1239, 2002.

Zhang Y, Ma C, Delohery T, Nasipak B, Foat BC, Bounoutas A, Bussemaker HJ, Kim SK, and Chalfie M. Identification of genes expressed in *C. elegans* touch receptor neurons. *Nature* 418: 331-335, 2002.

Figure Legends

Figure 1. KCNK gene protein features and OWEN alignments. A) Protein sequence of K01D12.4. Critical ion-selectivity motif (GXG) is in bold; upstream aromatic amino acid residue (F) is in bold and italics. Note that appropriate quantity of amino acids are present around GXG motifs to form four transmembrane alpha helices (see RESULTS). B) OWEN alignments to K01D12.4. Arrows show direction of transcription (5' – 3'). "Exons" track shows exon annotations from *C. elegans*. "*" denotes exon of gene K01D12.5 that flanks the 5' intergenic region of K01D12.4; "***" denotes the exon of gene K0D12.15 that flanks the 3' intergenic region. Note that 5' intergenic region is flanked by two 5' gene ends and 3' intergenic region is flanked by two 3' gene ends. "OWEN Exons" track shows *elegans-briggsae* alignments generated by OWEN that were within exons or overlapped exons; pieces that extended beyond exon were trimmed to within exon for clarity (other alignment pieces not shown for clarity). Note excellent coverage of K01D12.4 exons (overall: 96.3% of nucleotides) and K01D12.5 exon (*; actual coverage 95.6%) and poor coverage of K012D12.15 (**; 8.6% coverage). "OWEN CNSs" track. CNSs from OWEN alignments after removal of pieces in exons and after statistical filtering (see RESULTS). Bar shows scalebar for 2000 nt, which applies to all tracks. Note presence of intronic CNSs and "clumping" of intergenic CNSs.

Figure 2. Global CNS Statistics. For all graphs, circle is *elegans*, box is *briggsae*. A) Distribution of CNS lengths (nts). Note that gaps are not included in this length measurement. Bins are 5 nts. B) Distribution of similarity values for CNSs (Bins are 0.05). C) Distribution of selective constraint values for CNSs (Bins are 0.05). D) Distribution of transition/transversion ratio ("mutation ratio") (Bins are 0.10). For A-C, 476 total CNSs; for D, 459. Note excellent correlation between *briggsae* and *elegans* for A-C.

Figure 3. Spatial Patterns of CNSs. A) Clumping of CNSs. Histogram of inter-CNS distances (nts; 50 nt bins). Measurement of interCNS distance is from end of one CNS to start of next; distances that cross boundaries of exons are not included. Line is exponential fit. B) Histogram of CNS freq/nt in intergenic regions near ends. 1) first 300 nts of 3' intergenic region 2) first 300 nts of 5' intergenic region. Only intergenic regions >300 nts were included; n=28 regions for 3'; n=33 regions for 5'.

Figure 4. Differential GC content of CNS parts of ncDNA and non-CNS parts of ncDNA. For each region with CNSs, total GC content of CNSs and total GC content of non-CNSs were measured. A) Distributions of %GC content for CNSs and non-CNSs. solid circle is CNSs; open circle is non-CNSs. B) Distribution of paired difference in GC content, per region with CNSs, of CNS portions vs. non-CNSs portions. Note majority of regions have greater GC content in CNSs vs non-CNSs. Bins in A and B are 3%.

Figure 5. Intergenic Region CNSs. A) Length of intergenic region is correlated with number of 5' ends abutting region. Here, '0' denotes zero 5' ends (region is 3'-3' ends); '1' denotes 1 5' end (region is 3'-5' or 5'-3'); '2' denotes 2 5' ends (region is 5'-5'). B) Number of CNSs correlates with intergenic region length.

Figure 6. Intronic Region CNSs. A) Distribution of intronic CNSs/gene. Note that 13/33 have zero; 18/20 have 1-9 CNSs. B) Correlation of number of CNSs and summed length of introns in gene. Note if two large (>30 CNSs) points are removed, correlation still exists but is weakened (see RESULTS, data not shown). C) Introns possessing CNSs are distributed toward the ends of genes. A normalization routine was used to accommodate different numbers of introns/gene (see RESULTS and METHODS). Bins are 0.10. Values toward zero are toward 5' end; values toward 1 are toward 3' end.

Table 1: List Of 33 Accepted KCNK Genes From *C. elegans*

All of these genes passed criteria as described in RESULTS

These are contig names as found in <http://www.wormbase.org> (WS126)

F21C3.1
M110.2
ZK1067.5
B0334.2
F17C8.5
F22B7.7
ZC410.4
ZK1251.8
K04A8.4
F20A1.7
F29F11.4
R04F11.4
K01D12.4
C24A3.6
T06H11.1
C40C9.1
T01B4.1
F19D8.1
F55C5.3
M04B2.5
C33D12.3
C52B9.6
F36A2.4
Y47D3B.5
W06D12.5
K06B4.12
F31D4.7
R12G8.2
C48E7.9
F34D6.3
C24H11.8
T28A8.1
Y37A1B.11

Figure 1

A

MHKLQFARGIGQREMLRANTLPSITRAKVGCFARLRIYEENARFVLICII
LIVYLAFGAILFHWLEWENEVDERIAIDNRMADYQKVYCKHKPLNECDFE
EMVRFISDGATSGLLNSRSRFDHLGSLF**FSATVISTIGFGT**STPRTHLGR
FITIVYGVVGCTCCVLFFNLFLERLVTGMSYILRSLRERKIRYRLKESGN
KPVTLNNDNFNESSSSCGGHMDNWRPSVYKVFILFSMCLVLITASAG
IYSVVENWNYIDSLY**FCFISFATIGFGDY**VSNQQDVTRMSPDLYRFVNFC
LLTLGACFFYCLSNVSSIVVRQLLNWMIKKMDVKVEDRSFLCFKKRRRYM
GLGLRPPKGISNLLAGDNCCIEHLNVFRIRYDI

B

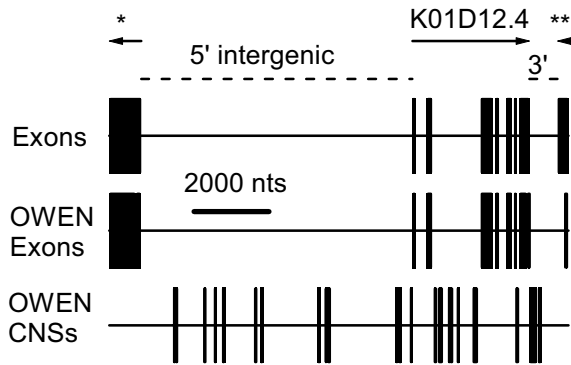


Figure 2

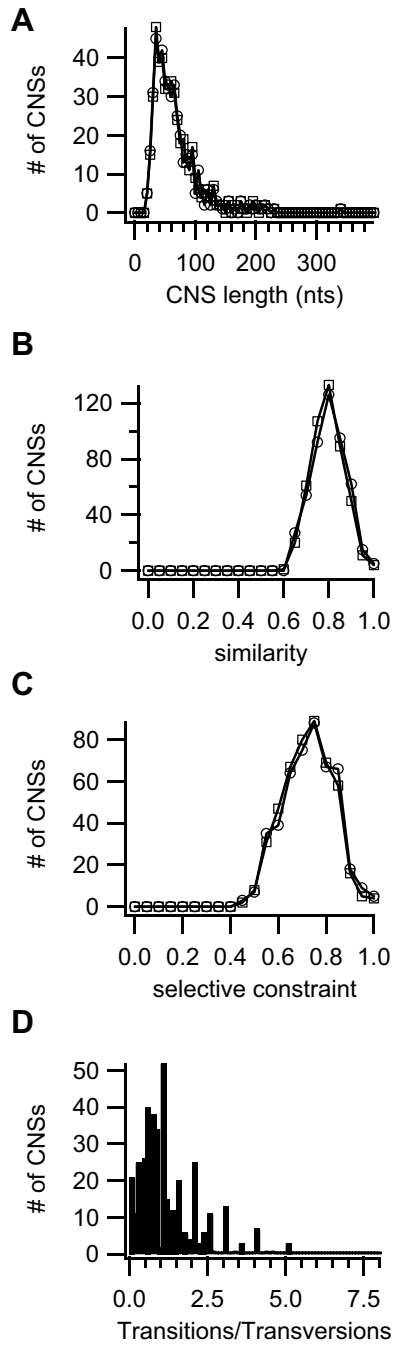


Figure 3

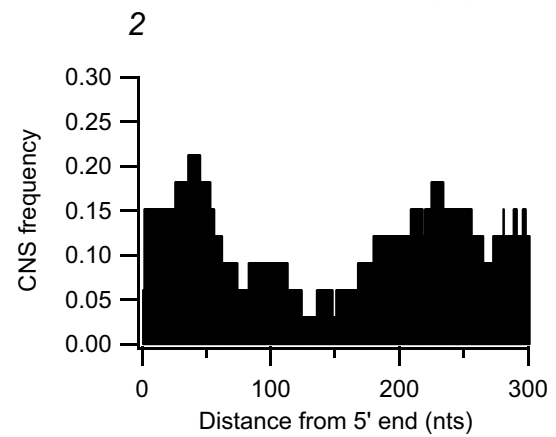
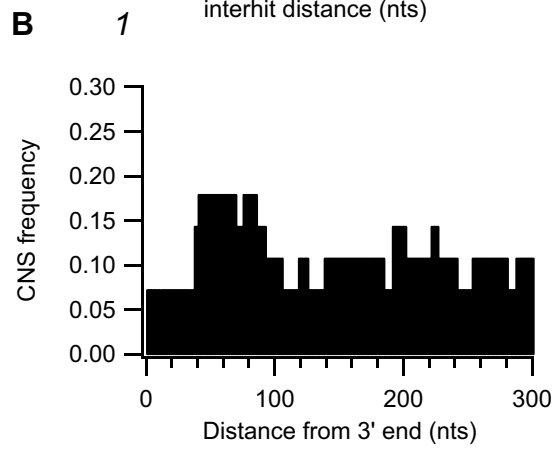
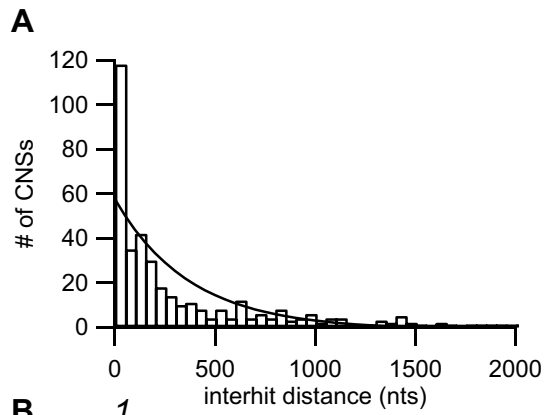


Figure 4

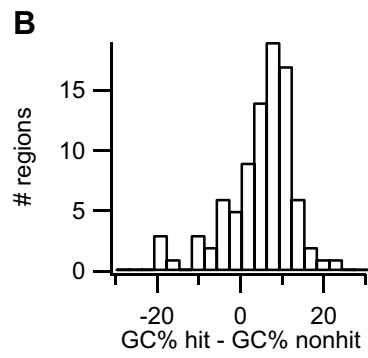
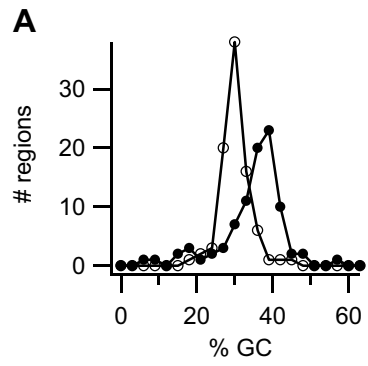
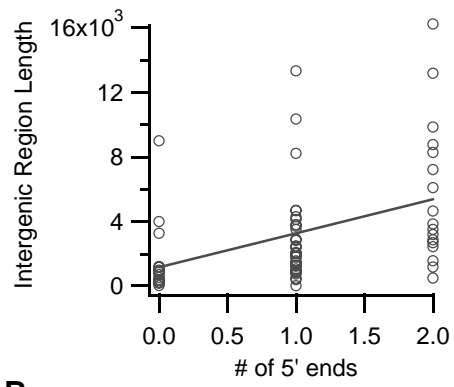


Figure 5

A



B

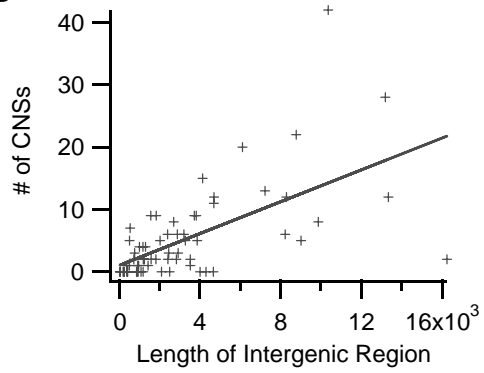
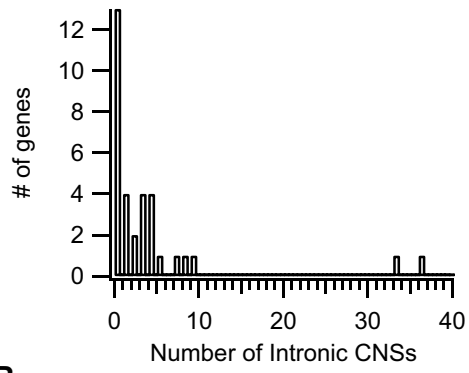
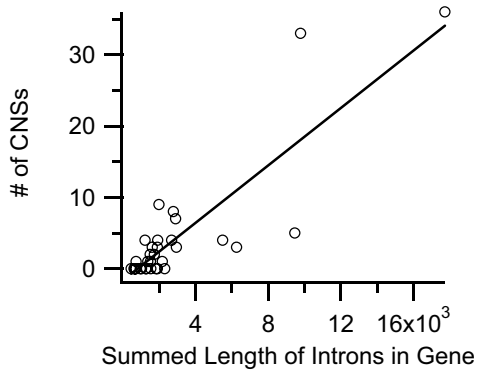


Figure 6

A



B



C

