Human Molecular Genetics, 2005, Vol. 14, No. 24 3837–3845
doi:10.1093/hmg/ddi408
Advance Access published on November 3, 2005

Evaluation of the *RET* regulatory landscape reveals the biological relevance of a HSCR-implicated enhancer

Elizabeth A. Grice[1], Erin S. Rochelle[1], Eric D. Green[3], Aravinda Chakravarti[1] and Andrew S. McCallion[1,2,*]

[1]McKusick-Nathans Institute of Genetic Medicine, [2]Department of Comparative Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA and [3]Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA

Received September 2, 2005; Revised and Accepted October 26, 2005 GenBank accession nos[†]

Evolutionary sequence conservation is now a relatively common approach for the prediction of functional DNA sequences. However, the fraction of conserved non-coding sequences with regulatory potential is still unknown. In this study, we focus on elucidating the regulatory landscape of *RET*, a crucial developmental gene within which we have recently identified a regulatory Hirschsprung disease (HSCR) susceptibility variant. We report a systematic examination of conserved non-coding sequences ($n = 45$) identified in a 220 kb interval encompassing *RET*. We demonstrate that most of these conserved elements are capable of enhancer or suppressor activity *in vitro*, and the majority of the elements exert cell type-dependent control. We show that discrete sequences within regulatory elements can bind nuclear protein in a cell type-dependent manner that is consistent with their identified *in vitro* regulatory control. Finally, we focused our attention on the enhancer implicated in HSCR to demonstrate that this element drives reporter expression in cell populations of the excretory system and central nervous system (CNS) and peripheral nervous system (PNS), consistent with expression of the endogenous RET protein. Importantly, this sequence also modulates expression in the enteric nervous system consistent with its proposed role in HSCR.

INTRODUCTION

The ability to impute function from analysis of DNA sequence alone remains an immense challenge in human genetics. Although protein-coding sequences can be predicted with relative ease, our understanding of the nature and identity of functional non-coding DNA is still rather limited. To a first approximation, functional DNA sequences can be predicted based upon evolutionary sequence conservation; functional regions are less tolerant of nucleotide substitution than non-functional (neutral) sequences (1), and thus evolve more slowly. Consistent with this hypothesis, coding sequences may be readily identified based on evolutionary conservation. Interestingly, of the 5% of the human genome that is estimated to be evolving more slowly than the neutral rate (2), less than one-third actually encodes protein. The remainder, conserved non-coding sequences, are commonly predicted to regulate temporal, spatial and quantitative aspects of gene expression (2,3), among other roles. However, unlike coding sequences, there is no vocabulary beyond conservation to guide the prediction of biological relevance of non-coding sequences (4). Consequently, there is significant interest in identifying and characterizing functional non-coding sequences.

The availability of an increasing number of vertebrate genome sequences, along with the development of bioinformatic analysis tools, has made the comparison of large

*To whom correspondence should be addressed at: McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, BRB Room 449, 733 N. Broadway, Baltimore, MD 21205, USA. Tel: +1 4432875624; Fax: +1 4106148600; Email: amccalli@jhmi.edu

[†]hg17, chr10: 42754810–43104810 (human); Mm3, chr6:118646816–119036816 (mouse); AC125509 and AC125512 (baboon); AC124166 (cat); AC138567 (chicken); RP43-171H18 (chimpanzee); AC124163 and AC124164 (cow); AC123973 (dog); AC124911 and AC125500 (fugu); AC122156 and AC124165 (pig); AC114881 (rat); AC135546 (tetraodon) and AC124155 (zebrafish)

genomic sequence intervals a feasible approach for the identification of putative regulatory sequences. However, despite increasing numbers of sequences identified through comparative sequence analysis, only a subset of conserved non-coding sequences identified at a handful of loci have been functionally characterized (5–13). The paucity of functional data for non-coding sequences represents a substantial impediment to evaluation of the potential role of non-coding variation in human disease. Although non-coding variation is predicted to play a significant role in common human disease (4,14,15), only ~1% of known human disease-associated mutations occur in regulatory sequences, localizing predominately within minimal promoter regions (16). Until recently, mutation detection in non-coding regions was almost exclusively restricted to sequences adjacent to the transcription start site; several classic examples of regulatory mutations have been identified in this way. Consequently, this number probably represents a gross underestimate of disease-causing non-coding mutations.

RET is a crucial developmental gene that encodes a receptor tyrosine kinase essential for normal embryonic development and neuronal maintenance. The protein is expressed throughout the CNS and PNS, and excretory system during embryogenesis, mediating signals influencing cell proliferation, differentiation, migration and apoptosis (17). *RET* is the major susceptibility gene in Hirschsprung disease (HSCR), a relatively common congenital disorder in which both non-coding and coding mutations are predicted to underlie disease susceptibility (14,18). We recently identified an enhancer sequence at *RET*, which contains a relatively common HSCR susceptibility mutation (15). Although the HSCR associated allele reduces *in vitro* enhancer activity 6-fold compared with the non-associated (wild-type) allele (15), the biological relevance of this enhancer sequence is unknown. Furthermore, the fraction of conserved non-coding sequences at this locus with regulatory potential is unknown. There are, in fact, few existing reports of comprehensive functional evaluation of conserved non-coding sequences of even a single locus, impeding attempts to examine association between non-coding variation and disease.

Here, we report a systematic examination of conserved non-coding sequences identified in a 220 kb interval encompassing *RET*. By employing a cell-based reporter assay, we demonstrate that most amplicons containing identified human *RET* conserved non-coding elements are capable of enhancer activity *in vitro*, and the majority of the elements exert cell type-dependent control. We also demonstrate that a selected subset of regulatory elements can bind nuclear protein in a cell type-dependent manner consistent with their *in vitro* activity. Importantly, the most striking neuronal enhancer identified in this study is the one we have implicated in HSCR based on association-genetic analysis in human patients (15). We report the *in vivo* function of this enhancer (MCS + 9.7) via transgenesis in mouse, demonstrating that it exerts regulatory control consistent with the endogenous RET protein. This enhancer drives reporter expression in cell populations of the excretory system, CNS and PNS and, specifically, in the digestive tract during embryogenesis in a manner consistent with its proposed role in HSCR.

RESULTS

Comparative sequence analysis identifies multi-species conserved sequences at *RET*

To identify conserved non-coding sequences at *RET*, we compared genomic sequence of an ~350 kb segment encompassing human *RET* (chr10: 42754810–43104810; UCSC hg17) with sequence from the orthologous intervals in 12 non-human vertebrates (chimpanzee, baboon, cow, pig, cat, dog, rat, mouse, chicken, zebrafish, *Fugu* and *Tetraodon*). The generation of genomic sequences used in this study has been described previously (15). Sequences were first aligned to the human reference sequence using AVID (19) and then visualized using the mVISTA tool (http://genome.lbl.gov/vista/index.shtml) (20,21) under established parameters of ≥70% identity, ≥100 bp (6). These criteria identified a total of 132 sequences conserved between the human reference and at least one other non-primate vertebrate. Forty-eight conserved sequences physically overlap exons of *RET* or other predicted genes. The remaining 84 conserved sequences are likely non-coding because no matching cDNA sequence or open-reading frame ≥20 amino acids in length were detected. Identification of these sequences was restricted to alignments between mammalian orthologs. All conserved sequence elements also overlapped predictions made by the method of Margulies *et al.* (22), a quantitative algorithm that uses phylogeny to calculate the probability of observing a given number of sequence identities at each base position. Although additional conserved sequences may be identified under modified criteria or through additional algorithms (23,24), the approaches we have used are both validated and established methods that clearly identify sequences evolving more slowly than the neutral rate (25).

In vitro analysis of *RET* multi-species conserved sequences

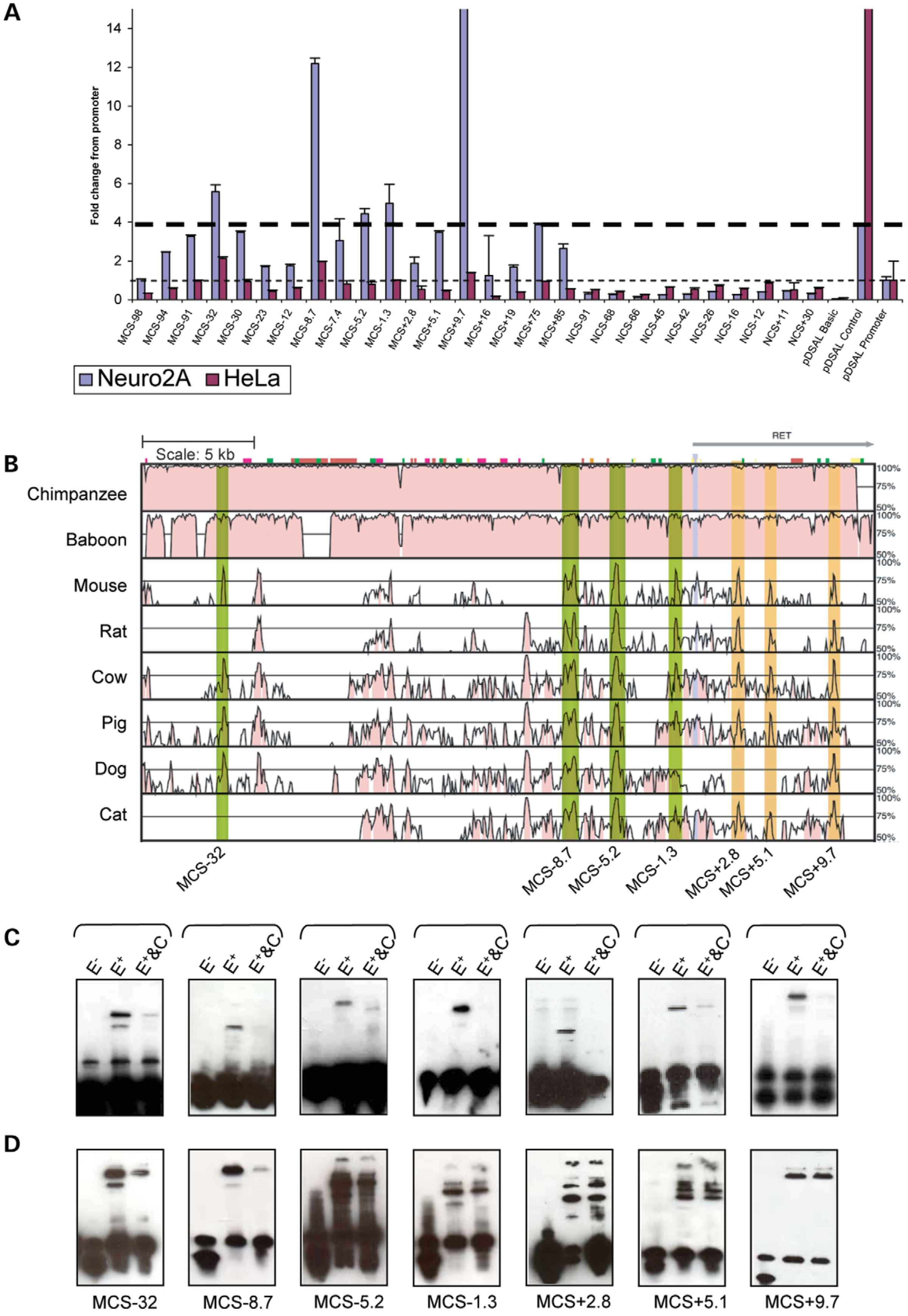
We directly tested the hypothesis that conserved non-coding sequences regulate gene expression by examining their potential to function as enhancers or repressors *in vitro*. To focus our efforts on conserved non-coding sequences present in multiple vertebrates, we prioritized a selection of those present in three or more non-primate mammals ($n = 45$ of 84 non-coding sequences). We refer to these as multi-species conserved sequences (MCSs). Because the boundaries of functional non-coding elements are not well defined, MCSs were frequently amplified in groups of two or more elements (Table 1); the average size of amplicons was 1.9 kb. We amplified 18 regions, encompassing all identified MCSs from human genomic DNA; primer sequences are listed in Supplementary Material, Table S1. These amplicons were subcloned in the context of the SV40 promoter and luciferase; completed constructs were termed pD*Sma_RET*_MCS*, where * denotes the distance (kb) and relative position (+ or −; 5′ or 3′, respectively) from the *RET* transcription start site. We selected two cell lines for these assays: the RET expressing neuroblastoma cell line, Neuro-2A, and the epithelial cell line, HeLa in which RET is not expressed. When transiently transfected into Neuro-2A cells, >80% (15/18) of MCS constructs (pD*Sma_RET*_MCS*) demonstrated increased luciferase reporter expression compared

Table 1. Description of non-coding amplicons encompassing identified MCSs

MCS construct	Amplicon coordinates	Amplicon size (bp)	Number of MCS in construct	MCS start	MCS end
MCS −98	42794674-42797159	2486 bp	6	42794705 42794862 42795069 42795971 42796639 42796905	42794799 42795036 42795453 42796066 42796743 42797109
MCS −94	42798448-42800055	1608 bp	2	42799407 42799550	42799537 42799692
MCS −91	42801327-42803770	2444 bp	2	42801825 42803295	42802059 42803650
MCS −32	42860115-42860427	414 bp	1	42860184	42860397
MCS −30	42862042-42863343	1302 bp	3	42862433 42863012 42863177	42862883 42863156 42863275
MCS −23	42869519-42872891	3373 bp	4	42869985 42870538 42870688 42871537	42870259 42870657 42871284 42871985
MCS −12	42880822-42883700	2879 bp	3	42880899 42882148 42882378	42881337 42882296 42882844
MCS −8.7	42883576-42884765	1190 bp	2	42883670 42884020	42883888 42884443
MCS −7.4	42884744-42886753	2010 bp	2	42885043 42886396	42885188 42886523
MCS −5.2	42886009-42887780	1772 bp	1	42887074	42887422
MCS −1.3	42891094-42891708	615 bp	1	42891278	42891461
MCS +2.8	42893765-42896916	3642 bp	1	42895462	42895726
MCS +5.1	42897555-42898617	1063 bp	1	42897720	42897886
MCS +9.7	42901818-42902717	900 bp	1	42902073	42902291
MCS +16	42908045-42910632	2588 bp	4	42908150 42908343 42908455 42909539	42908289 42908438 42908727 42909860
MCS +19	42910460-42912656	2197 bp	4	42910485 42911525 42911912 42912292	42910579 42911669 42912107 42912565
MCS +75	42966689-42969556	2868 bp	4	42967623 42968035 42968825 42969395	42967727 42968132 42968921 42969508
MCS +85	42977372-42979276	1905 bp	3	42977407 42977946 42979089	42977680 42978103 42979228

Sequences correspond to positions within the May 2004 (UCSC hg 17) build.

with a control vector (pD*Sma*_promoter) in which luciferase expression was driven by the SV40 promoter fragment alone (Fig. 1A). These results suggest that the majority of non-coding *RET* MCS amplicons may play a role in regulating gene expression. We next determined whether such regulatory activity was consistent with tissue-dependent *RET* regulatory control. We directly examined this question by conducting these assays using the HeLa cell line. Consistent with the absence of RET expression in HeLa cells, <17% (3/18) of MCS containing constructs (pD*Sma*_*RET*_MCS*) demonstrated luciferase expression that was greater than the control pD*Sma*_promoter construct. However, 67% (12/18) of MCS constructs appeared to actively repress luciferase expression in HeLa cells. All assays were conducted in triplicate and were consistent upon repetition. These data are consistent with the tissue-dependent nature of *RET* regulation. Importantly, we likewise examined the regulatory potential of constructs containing non-conserved sequences (NCSs) ($n = 10$). These sequences failed to drive luciferase expression at levels significantly greater than the control (pD*Sma*_promoter; Fig. 1A)


A
Fold change from promoter
0
2
4
6
8
10
12
14
MCS-98
MCS-94
MCS-91
MCS-32
MCS-30
MCS-23
MCS-12
MCS-8.7
MCS-7.4
MCS-5.2
MCS-1.3
MCS+2.8
MCS+5.1
MCS+9.7
MCS+16
MCS+19
MCS+75
MCS+85
NCS-91
NCS-88
NCS-66
NCS-45
NCS-42
NCS-26
NCS-16
NCS-12
NCS+11
NCS+30
pDSAL Basic
pDSAL Control
pDSAL Promoter
Neuro2A
HeLa
B
Scale: 5 kb
RET
Chimpanzee
Baboon
Mouse
Rat
Cow
Pig
Dog
Cat
100%
75%
50%
MCS-32
MCS-8.7
MCS-5.2
MCS-1.3
MCS+2.8
MCS+5.1
MCS+9.7
C
E-
E+
E+&C
D
MCS-32
MCS-8.7
MCS-5.2
MCS-1.3
MCS+2.8
MCS+5.1
MCS+9.7

in either of the above cell types. NCS sequences, tabulated in Supplementary Material, Table S2, were distributed evenly throughout the 220 kb interval and ranged in size from 0.5 to 1.6 kb.

Identified MCSs demonstrate sequence-specific binding of nuclear protein

Regulatory sequences commonly mediate their effect upon binding transcription factors that instruct their regulatory control. Thus, we hypothesized that discrete sequences within regulatory MCSs would bind nuclear protein in a sequence-specific and cell type-dependent manner consistent with their behavior in the above described luciferase assay. To test this postulate, we conducted electrophoretic mobility shift assays (EMSAs) on selected sequences within a subset of MCSs. First, we prioritized MCS amplicons demonstrating the greatest magnitude of effect in luciferase assays performed in neuronal (Neuro-2A) cells, under the assumption that *in vitro* activity provides a reasonable surrogate for *in vivo* functional potential. Specifically, we selected MCSs capable of driving luciferase expression at levels $\geq$4-fold than the promoter only construct ($n = 5$; MCS −32; MCS −8.7; MCS −5.2; MCS −1.3; MCS +9.7). These sequences are highlighted (green) in Figure 1B. Secondly, we additionally selected all MCSs localizing to the genetic interval previously implicated in HSCR susceptibility (*RET* intron 1; $n = 3$) (14,15,26–28), under the assumption that one or more of these MCSs may be relevant to HSCR. These sequences (MCS +2.8; MCS +5.1; MCS +9.7) are highlighted in orange in Figure 1B.

Conserved non-coding sequences are reported to be enriched for functional transcription factor binding sites (TFBSs) (9). We examined MCSs within selected amplicons for known TFBS using TESS (Transcription Element Search Site, URL: http://www.cbil.upenn.edu/tess) as described in Materials and Methods. Identified TFBSs are tabulated in Supplementary Material, Table S3. In light of these analyses, we synthesized oligonucleotides (30–50mer, sequences tabulated in Supplementary Material, Table S4) to include complete consensus sequences for predicted TFBS and examined their potential to bind nuclear protein using EMSAs. Protein binding was determined to be sequence-specific if excess unlabeled oligonucleotide was able to displace labeled oligonucleotide (E$^+$ & C lanes, Fig. 1C or D). 17/25 oligonucleotides, corresponding to 7/7 MCS amplicons demonstrated sequence-specific binding to Neuro-2A nuclear protein (Fig. 1C). Only 5/25 oligonucleotides, corresponding to 2/7 MCS amplicons, demonstrated sequence-specific binding of HeLa nuclear protein (Fig. 1D). The majority of the oligonucleotides readily bound nuclear extract from HeLa cells; however, as the bound protein was not significantly displaced by excess unlabeled oligonucleotide, the binding was not sequence-specific, but is likely to be an artifact of the assay. Additionally, multiple shifting bands were frequently observed when oligonucleotides were incubated with HeLa nuclear protein. These multiple bands likely represent oligonucleotides bound by full or partial protein complexes. However, these patterns were not consistent with sequence-specific interactions, as they were not competed away by excess probe. Importantly, these data are also consistent with the regulatory control of gene expression observed, for the corresponding MCS amplicons (MCS −32; MCS −8.7), in the above luciferase assays, as summarized in Table 2. MCS −32 and MCS −8.7 were the only two MCSs capable of enhancing luciferase expression in HeLa cells, and they were the only two MCSs that bound nuclear extract from HeLa cells in a sequence-specific manner.

Table 2. *RET* MCSs exert cell type-dependent control

	Luciferase assay		EMSA	
	Neuro2A	HeLa	Neuro2A	HeLa
MCS −32	+	+	3/3	3/3
MCS −8.7	+	+	3/5	2/5
MCS −5.2	+	−	2/4	0/4
MCS −1.3	+	−	2/2	0/2
MCS +2.8	+	−	2/3	0/3
MCS +5.1	+	−	2/2	0/2
MCS +9.7	+	−	4/6	0/6

MCSs capable of driving luciferase expression in the indicated cell-types, significantly greater than the promoter alone, are designated with (+). Proportion of oligonucleotides within regulatory MCSs that can bind nuclear protein in sequence specific manner by EMSA are designated.

Consistent with all of the above postulates, the recently identified HSCR-susceptibility variant localizes to enhancer sequence MCS +9.7, which demonstrates the greatest magnitude of effect of all examined MCSs, driving luciferase (Fig. 1A). Furthermore, the identified variant lies within an examined oligonucleotide sequence demonstrating cell type-dependent protein binding. However, although the HSCR-susceptibility variant lies within a predicted SRF-binding site (Supplementary Material, Table S3) and within one nucleotide of a predicted retinoic acid receptor (RARα1, RARβ, RARγ) site, a potential role for these sites remains unclear in the absence of a full understanding of the biological

Figure 1. *In vitro* characterization of conserved non-coding sequences at *RET*. (**A**) Luciferase expression of 45 MCSs contained within 18 amplicons. pD*SMA*_MCS* constructs were analyzed in Neuro-2A (blue bars) and HeLa (red bars) cell lines. Expression values are normalized against promoter only construct (pD*SMA*_promoter) expression (thin dotted line). Likewise, 10 NCSs were analyzed in the same assay. Additional controls included pDSMA_basic and pDSMA_control constructs; the former contained a luciferase ORF in the absence of a promoter, and the latter comprised SV40 promoter and enhancer sequences in combination with a luciferase ORF. All assays were conducted in triplicate and consistent upon replication; error bars report standard error in each instance. (**B**) mVISTA plot comparing human reference sequence with orthologous sequence from 8 mammals (window shown = 50 kb/350 kb). Green-highlighted regions designate *RET* MCS sequences that were capable of driving luciferase expression 4-fold compared with the promoter alone (see thick dotted line in Fig. 1A). Additionally, orange-highlighted regions designate *RET* MCS sequences localized to intron 1, wherein peak transmission distortion (by TDT) occurs in HSCR (15). Colored peaks indicate sequence conservation of $\geq$70% identity and $\geq$100 nucleotides. Red, non-coding; Blue, *RET* exons. (**C**) EMSAs using Neuro-2A cell extract with 30–50mer oligonucleotides located within *RET* MCSs. E−, no extract; E$^+$, extract added; E$^+$ & C, extract and competing unlabeled oligonucleotide added. (**D**) Corresponding EMSAs were also performed for each oligonucleotide using HeLa nuclear extract.

relevance of MCS +9.7. For this, as for any identified regulatory MCSs, such an understanding ultimately necessitates their analysis *in vivo*.

Transgenic analysis of MCS +9.7 regulatory control

To begin to determine the *in vivo* role of regulatory MCSs identified at this locus, we prioritized MCS +9.7, having previously demonstrated significant association between HSCR susceptibility and an MCS +9.7 variant (15). To test MCS +9.7 for its ability to spatially and temporally modulate gene expression, we subcloned the MCS +9.7 amplicon into a β-galactosidase (LacZ) reporter vector in the context of the mouse heat shock protein 68 (*hsp68*) promoter. The transgenic construct was injected into fertilized mouse oocytes, and multiple stable transgenic lines (G0) were identified ($n = 4$). We then established timed matings to facilitate examination of the resulting G1 embryos at time points overlapping the critical period of RET activity during embryogenesis (10.5–14.5 dpc, days *post coitum*).

MCS +9.7 drives LacZ reporter expression in a manner consistent with many aspects of the temporal and spatial expression of RET (17,29–31). Most notably, LacZ expression is detected within the external gut loops at 12.5 dpc (Fig. 2A) consistent with RET expression in the enteric nervous system during embryogenesis (Fig. 2B) and its proposed role in HSCR (15). By 14.5 dpc, reporter signal is detected throughout the length of the gut consistent with RET expression during the colonization of the gut by neural crest-derived enteric ganglia (29,31). Reporter expression is also detected in the developing sensory and autonomic ganglia of the trunk of MCS +9.7 transgenic embryos at all time points observed (10.5–14.5 dpc). At 10.5 dpc, faint and diffuse LacZ signal was detected in the spinal cord in positions consistent with truncal neural crest émigrés immediately prior to their condensation to form the dorsal root ganglia (DRG) (data not shown) (32). At later time points (12.5 dpc–14.5 dpc), LacZ staining intensified, becoming punctuate (Fig. 2C and E), consistent with RET expression in DRG (Fig. 2D). The identity of this cell population was confirmed in a transverse section through the trunk of a 12.5 dpc transgenic embryo which illustrates localized staining in the DRG population. This expression pattern is also accompanied by reporter signal in the more ventral portion of the spinal cord, consistent with *RET* expression in the motor neuron column (Fig. 2F). This ventrally localized domain of staining extends along the entire anterior–posterior axis of the spinal cord and into the hindbrain (data not shown). Additionally, MCS +9.7 transgenic embryos displayed LacZ staining in the brain. Staining was localized to the midbrain, the pons and the forebrain (Fig. 2C, G and I). Indicated in

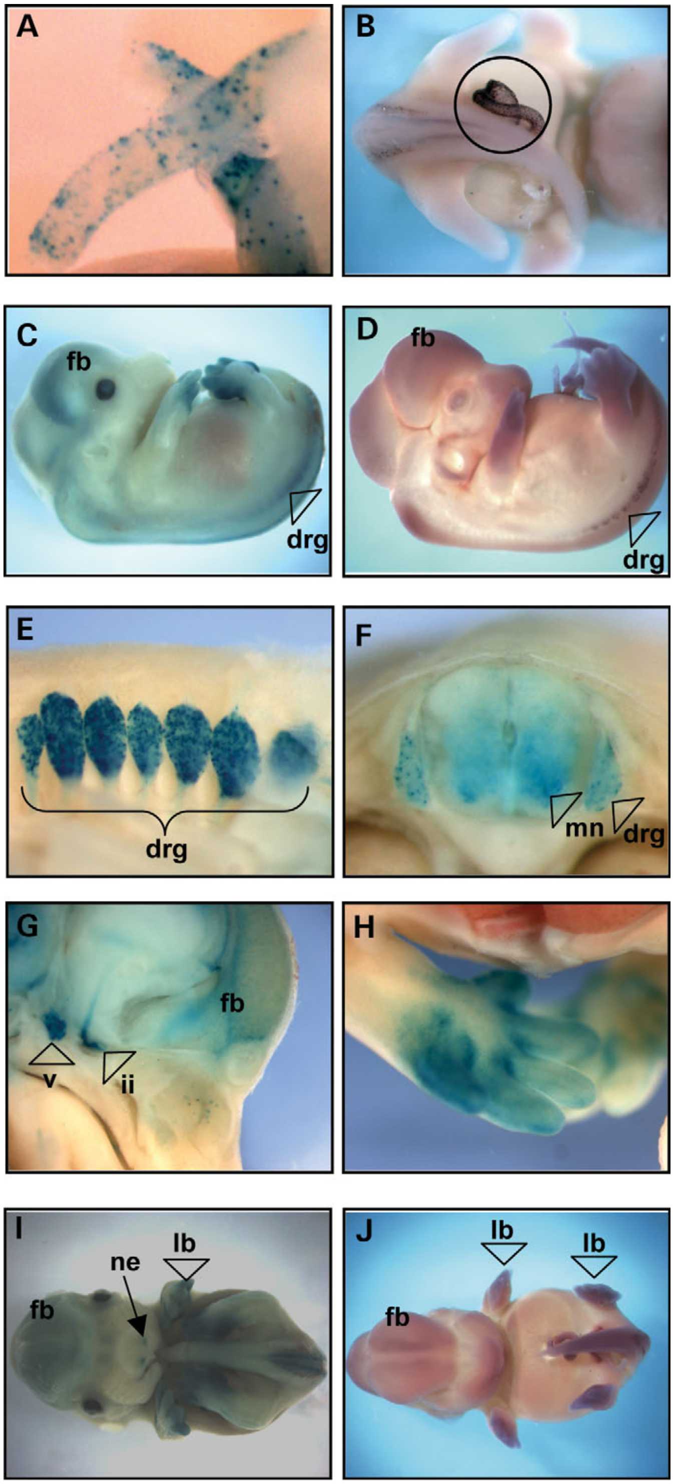


Figure 2. MCS + 9.7 demonstrates regulatory control consistent with RET expression. (**A**) LacZ expression in the external gut loop of a 12.5 dpc embryo. (**B**) Expression of *Ret* in the external gut is indicated by *in situ* hybridization to *Ret* antisense probe. (**C**) LacZ staining in forebrain and DRG is indicated in a 12.5 dpc whole-mount embryo. (**D**) Expression of *ret* in whole mount 12.5 dpc embryo as detected by *in situ* hybridization. Forebrain and DRG are indicated. (**E**) Medial view of LacZ expression in the thoracic DRG sagitally bisected in 14.5 dpc embryo. (**F**) Transversally bisected view of stained thoracic motorneuron columns and DRG in 14.5 dpc embryo. (**G**) Medial view of 14.5 dpc embryo head sagitally bisected. Prominent staining corresponds to trigeminal ganglia (V) and optic ganglia (II). Staining is also indicated in forebrain. (**H**) Limb staining in 14.5 dpc embryo. Strongest LacZ espression is localized to the mesenchyme between the digits. (**I**) Ventral view of 12.5 dpc whole-mount embryo. Staining in forebrain, limb and neural epithelium is indicated. (**J**) Corresponding *in situ* hybridization with *ret* antisense probe. fb, forebrain; drg, dorsal root ganglia; lb, limb; ne, nasal epithelium; g, gut; mn, motorneurons; v, trigeminal cranial ganglia (V); ii, optic cranial ganglia (II).

Figure 2G is LacZ staining in the trigeminal ganglia (V) and the optic ganglia (II). This staining pattern is consistent with the reported role for *RET* in the development of all cranial ganglia (29). LacZ staining is also detected in the cells of the nasal epithelium (Fig. 2I) (31) as well as in the forelimbs and hind limbs of MCS +9.7 transgenic embryos beginning at 12.5 dpc; staining in the limbs is predominantly localized to the mesenchyme between the digits (Fig. 2H), consistent with expression of the RET ligands GFRα1 and GFRα2 (31). These data were consistent among multiple (3/4) examined transgenic lines.

DISCUSSION

Until recently, sequences have been broadly categorized as genic or non-genic and coding or non-coding. However, these definitions are inadequate to provide complete functional annotation of genomes. Comparisons of orthologous sequences between genomes have already uncovered regions of predicted functional constraint. Such regions comprise known genes, previously unknown protein-coding genes, regulatory RNAs and non-coding regulatory sequences. Recent analysis of the complete human and mouse genomes demonstrated that 40% of their sequences could be aligned at the nucleotide level but only 5% appeared to be under active selection and therefore predicted to be functional (2). Notably, conserved non-coding elements comprise more than twice as much of the human genome as protein coding sequences (2,8). Importantly, the vast numbers of conserved non-coding sequences makes comprehensive determination of their function particularly challenging. An, as yet, undetermined fraction of these sequences contribute to control of temporal, spatial and quantitative aspects of gene expression (3). Despite their predicted role in common inherited human disorders, our ability to associate non-coding variation with disease is hampered by an incomplete understanding of the identity and composition of regulatory sequences. Systematic functional evaluation of conserved non-coding sequences will represent a significant step towards understanding this role.

Consistent with previous reports, we demonstrate that sequences conserved in multiple mammals are frequently regulatory (6,10,33,34). Our data suggest that most MCS amplicons examined at *RET* function as enhancers often in a tissue-dependent manner. The vast majority of examined sequences enhanced luciferase expression in neuronal cells (Neuro-2A) but not in epidermal cells (HeLa), consistent with RET expression. Furthermore, our data suggest that a single MCS amplicon may function as an enhancer in one cell type, yet repress expression in another. By selecting informative cell types, we demonstrate that both enhancer and suppressor functions may be readily discerned *in vitro*. Although these assays may miss many of the subtleties of *in vivo* function, their utility in determining tissue-dependent behavior of examined sequences is clear. However, examination of the biological or disease relevance of any regulatory MCS ultimately necessitates its evaluation *in vivo*.

MCS +9.7: LacZ transgenic mouse lines generated in this study display regulatory control of reporter expression in the PNS, CNS and excretory systems, and in the limb, consistent with the endogenous *Ret* gene. Furthermore, we demonstrate that MCS +9.7 is independently capable of regulating many aspects of *RET*-like expression; specifically, this element drives expression in the ENS, consistent with RET expression patterns and with a predicted role for a mutation in this sequence in HSCR susceptibility. However, given the number of regulatory MCS amplicons at this locus and the fine spatial and temporal regulation of *RET* in discrete cell subpopulations, we predict that many other sequences at *RET* also play complementary and/or cooperative regulatory roles. Critically, we have now demonstrated the potential disease relevance of MCS +9.7 through human genetic, *in vitro* and *in vivo* analyses. We are presently undertaking experiments to fully evaluate the biological impact of the HSCR-associated variant identified therein.

There is a well-recognized need for rapid functional screens of non-coding sequences. Importantly, although several recent reports have demonstrated that sequence comparisons at the vertebrate extremes provide a powerful filter for functional sequences, such ultraconserved elements are not abundant in vertebrate genomes and may be found in physical proximity to <1% of human genes ($n = 156$) (35). This suggests that most vertebrate regulatory sequences may not be detectably conserved across large evolutionary distances (humans to teleosts). Rather, comparison of sequence orthologs from multiple more closely related species (mammals only/teleosts only) may prove to be a more sensitive approach for identifying vertebrate regulatory sequences (22). Consistent with this prediction, our data suggest that sequences identified in this way are also frequently functional. Furthermore, identified MCSs may be evaluated in combination, and critical sequences therein may subsequently be dissected.

Our data demonstrate the utility of examining multiple MCSs in combination in order to decrease the numbers of analyses required to prioritize sequences for subsequent molecular and *in vivo* investigation. These data suggest that most amplicons encompassing conserved non-coding sequences that are identified under established criteria are frequently regulatory. Furthermore, tissue-dependent regulatory control may be inferred from *in vitro* analysis in appropriate cell lines, and candidate TFBS may be identified by molecular investigation. However, it should be noted that the activity of an MCS amplicon may reflect the activity of a single MCS or the additive or synergistic effects of several MCSs therein. Thus, further functional evaluation of MCS amplicons will be necessary to identify specific sequences or motifs responsible for their regulatory and/or disease potential. In summary, we demonstrate the power of combining *in silico*, *in vitro* and molecular analysis for the identification of regulatory sequences at any locus, ultimately to determine the biological and/or disease relevance of selected sequences *in vivo*. Critically, these data suggest that regulatory non-coding sequences are not restricted to those conserved at the vertebrate extremes.

MATERIALS AND METHODS

Generation of luciferase reporter constructs

Eighteen MCS regions, encompassing a total of 45 MCSs were PCR amplified from human genomic DNA using PCR primers incorporating Gateway® *attB* sequences (sequences specified in Supplementary Material, Table S1). Each MCS amplicon was cloned into pDONR221™, a Gateway entry vector, per manufacturer's protocol. Amplicons were then subcloned into a *Sma*I site in a Gateway modified pGL3 (Promega, Madison WI, USA) firefly luciferase vector containing an SV40 promoter and complete firefly luciferase open-reading frame (pD*Sma*). Plasmids containing only the SV40 promoter driving luciferase (pD*Sma*_promoter), the SV40 promoter and enhancer driving luciferase (pD*Sma*_control) and containing only the luciferase ORF (pD*Sma*_basic) served as experimental control vectors.

Transfection of reporter constructs

Transient transfections were performed using neuroblastoma (Neuro-2A, ATCC no. CCL-131) and HeLa cell lines (ATCC no. CRL-13011). Cell lines were cultured according to ATCC protocols (http://www.atcc.org). Approximately 10^5 cells were co-transfected (Lipofectamine Plus™, Invitrogen) with 0.4 μg of the appropriate pD*Sma* firefly luciferase plasmid (pD*Sma*_promoter, pD*Sma*_control or pD*Sma*_*RET*_MCS*) and 0.01 μg phRL-SV40 control renilla luciferase plasmid. Dual Luciferase® assays (Promega) were performed in accordance with manufacturer's instructions. Luciferase activity was assayed 24 h after transfection (Victor^{3TM} plate reader, Perkin Elmer; Monolight® 2010, Analytical Luminescence Laboratories, CA, USA). All assays were conducted in triplicate and were consistent upon repetition. Relative luciferase units (RLU) were calculated for each transfection and fold change from pD*Sma*_promoter RLU was estimated. Fold change values of each construct are reported with corresponding standard errors (Fig. 1A).

TFBS identification

MCS sequences were queried for TFBSs using TESS (Transcription Element Search Site, URL: http://www.cbil.upenn.edu/tess). TRANSFAC 4.0 strings were searched and all results with a maximum allowable string mismatch (t_{mm}) of 10%, a minimum log-likelihood ratio score ($t_{s\text{-}a}$) >14.0 and a minimum string length (t_w) >6 were considered. Filters were applied to restrict predictions to mammalian species.

Electrophoretic mobility shift assay

Nuclear proteins were extracted from Neuro-2A and HeLa cells using NE-PER® Nuclear and Cytoplasmic Extraction Kit (Pierce Biotechnologies, Rockford IL, USA). Oligos (30–50mer) within MCSs were designed based on level of conservation and relevant TFBSs present. Oligos were end labeled with biotin-ddUTP (Pierce Biotechnology) and annealed per manufacturer's protocol. About 4 fmol of labeled oligo was incubated for 20 min at room temperature with 1 × binding buffer, 50 ng/μl poly dI–dC and 4 μl Neuro-2A nuclear extract (5 μl for HeLa nuclear extract). About 20 pmol (5000-fold) unlabeled oligo was added to competition reactions. Gel shifts were detected using the LightShift® Chemiluminescent EMSA kit (Pierce Biotechnology) after transfer from 15% acrylamide gel to nylon membrane.

Mouse transgenic reporter assay

All animal studies were performed under protocols approved by the Johns Hopkins University Animal Care and Use Committee. MCS amplicons were subcloned into the Gateway ready vector phsp68/LacZ (a kind gift of Dr E.M. Rubin, LBNL). The constructs were purified and injected into mouse pronuclei by the Johns Hopkins University Transgenic Core. G0 mice were genotyped by PCR with LacZ-specific primers and MCS-specific primers. Positive G0 and F1 males were mated to CD1 females; 12:00 p.m. of the day that vaginal plugs were observed was defined as 0.5 dpc. Embryos were harvested at specified time points in cold PBS. Transgenic embryos were identified by PCR of yolk sac DNA using LacZ primers (forward primer: TTT CCA TGT TGC CAC TCG C, reverse primer: AAC GGC TTG CCG TTC AGC A). Transgenic embryos were assayed for β-gal using 5-bromo-4-chloro-3-indolyl-β-D-galactoside (Ultrapure™ X-gal; Sigma) as described (36).

In situ hybridization

Wild-type mouse embryos were harvested from timed matings established with CD1 mice. Non-radioactive whole-mount *in situ* hybridization was performed as described (37). Digoxigenin-labeled antisense probes were made with template 2.5 kb Ret (pmcRet7 NotI T7 RNA polymerase).

SUPPLEMENTARY MATERIAL

Supplementary Material is available at HMG Online.

ACKNOWLEDGEMENT

This study was supported by a grant from the National Institute of Child Health and Development and by a Basil O'Conner starter scholar award from the March of Dimes.

Conflict of Interest statement. None declared.

REFERENCES

1. Kimura, M. and Ota, T. (1971) On the rate of molecular evolution. *J. Mol. Evol.*, **1**, 1–17.
2. Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P. *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.
3. Pennacchio, L.A. and Rubin, E.M. (2001) Genomic strategies to identify mammalian regulatory sequences. *Nat. Rev. Genet.*, **2**, 100–109.
4. Pastinen, T. and Hudson, T.J. (2004) Cis-acting regulatory variation in the human genome. *Science*, **306**, 647–650.
5. Oeltjen, J.C., Malley, T.M., Muzny, D.M., Miller, W., Gibbs, R.A. and Belmont, J.W. (1997) Large-scale comparative sequence analysis of the human and murine Bruton's tyrosine kinase loci reveals conserved regulatory domains. *Genome Res.*, **7**, 315–329.
6. Loots, G.G., Locksley, R.M., Blankespoor, C.M., Wang, Z.E., Miller, W., Rubin, E.M. and Frazer, K.A. (2000) Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science*, **288**, 136–140.

7. Pennacchio, L.A., Olivier, M., Hubacek, J.A., Cohen, J.C., Cox, D.R., Fruchart, J.C., Krauss, R.M. and Rubin, E.M. (2001) An apolipoprotein influencing triglycerides in humans and mice revealed by comparative sequencing. *Science*, **294**, 169–173.
8. Thomas, J.W., Touchman, J.W., Blakesley, R.W., Bouffard, G.G., Beckstrom-Sternberg, S.M., Margulies, E.H., Blanchette, M., Siepel, A.C., Thomas, P.J., McDowell, J.C. *et al.* (2003) Comparative analyses of multi-species sequences from targeted genomic regions. *Nature*, **424**, 788–793.
9. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. and Lander, E.S. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, **423**, 241–254.
10. Frazer, K.A., Tao, H., Osoegawa, K., de Jong, P.J., Chen, X., Doherty, M.F. and Cox, D.R. (2004) Noncoding sequences conserved in a limited number of mammals in the SIM2 interval are frequently functional. *Genome Res.*, **14**, 367–372.
11. Pribnow, D. (1975) Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter. *Proc. Natl Acad. Sci. USA*, **72**, 784–788.
12. Emorine, L., Kuehl, M., Weir, L., Leder, P. and Max, E.E. (1983) A conserved sequence in the immunoglobulin J kappa–C kappa intron: possible enhancer element. *Nature*, **304**, 447–449.
13. Woolfe, A., Goodson, M., Goode, D.K., Snell, P., McEwen, G.K., Vavouri, T., Smith, S.F., North, P., Callaway, H., Kelly, K. *et al.* (2005) Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.*, **3**, e7.
14. McCallion, A.S., Emison, E.S., Kashuk, C.S., Bush, R.T., Kenton, M., Carrasquillo, M.M., Jones, K.W., Kennedy, G.C., Portnoy, M.E., Green, E.D. *et al.* (2003) Genomic variation in multigenic traits: Hirschsprung disease. *Cold Spring Harb. Symp. Quant. Biol.*, **68**, 373–381.
15. Emison, E.S., McCallion, A.S., Kashuk, C.S., Bush, R.T., Grice, E., Lin, S., Portnoy, M.E., Cutler, D.J., Green, E.D. and Chakravarti, A. (2005) A common sex-dependent mutation in a RET enhancer underlies Hirschsprung disease risk. *Nature*, **434**, 857–863.
16. Stenson, P.D., Ball, E.V., Mort, M., Phillips, A.D., Shiel, J.A., Thomas, N.S., Abeysinghe, S., Krawczak, M. and Cooper, D.N. (2003) Human Gene Mutation Database (HGMD): 2003 update. *Hum. Mutat.*, **21**, 577–581.
17. McCallion, A.S. and Chakravarti, A. (2004) In Epstein, C., Erickson, R. and Wynshaw-Boris, A. (eds), *Inborn Errors of Development*. Oxford University Press, San Francisco, Vol. 23, pp. 335–338.
18. Carrasquillo, M.M., McCallion, A.S., Puffenberger, E.G., Kashuk, C.S., Nouri, N. and Chakravarti, A. (2002) Genome-wide association study and mouse model identify interaction between RET and EDNRB pathways in Hirschsprung disease. *Nat. Genet.*, **32**, 237–244.
19. Bray, N., Dubchak, I. and Pachter, L. (2003) AVID: a global alignment program. *Genome Res.*, **13**, 97–102.
20. Frazer, K.A., Pachter, L., Poliakov, A., Rubin, E.M. and Dubchak, I. (2004) VISTA: computational tools for comparative genomics. *Nucleic Acids Res.*, **32**, W273–W279.
21. Mayor, C., Brudno, M., Schwartz, J.R., Poliakov, A., Rubin, E.M., Frazer, K.A., Pachter, L.S. and Dubchak, I. (2000) VISTA: visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics*, **16**, 1046–1047.
22. Margulies, E.H., Blanchette, M., Haussler, D. and Green, E.D. (2003) Identification and characterization of multi-species conserved sequences. *Genome Res.*, **13**, 2507–2518.
23. Cooper, G.M., Brudno, M., Green, E.D., Batzoglou, S. and Sidow, A. (2003) Quantitative estimates of sequence divergence for comparative analyses of mammalian genomes. *Genome Res.*, **13**, 813–820.
24. Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.
25. Dermitzakis, E.T., Reymond, A. and Antonarakis, S.E. (2005) Conserved non-genic sequences—an unexpected feature of mammalian genomes. *Nat. Rev. Genet.*, **6**, 151–157.
26. Griseri, P., Bachetti, T., Puppo, F., Lantieri, F., Ravazzolo, R., Devoto, M. and Ceccherini, I. (2005) A common haplotype at the 5′ end of the RET proto-oncogene, overrepresented in Hirschsprung patients, is associated with reduced gene expression. *Hum. Mutat.*, **25**, 189–195.
27. Pelet, A., de Pontual, L., Clement-Ziza, M., Salomon, R., Mugnier, C., Matsuda, F., Lathrop, M., Munnich, A., Feingold, J., Lyonnet, S. *et al.* (2005) Homozygosity for a frequent and weakly penetrant predisposing allele at the RET locus in sporadic Hirschsprung disease. *J. Med. Genet.*, **42**, e18.
28. Burzynski, G.M., Nolte, I.M., Bronda, A., Bos, K.K., Osinga, J., Plaza Menacho, I., Twigt, B., Maas, S., Brooks, A.S., Verheij, J.B. *et al.* (2005) Identifying candidate Hirschsprung disease-associated RET variants. *Am. J. Hum. Genet.*, **76**, 850–858.
29. Pachnis, V., Mankoo, B. and Costantini, F. (1993) Expression of the c-ret proto-oncogene during mouse embryogenesis. *Development*, **119**, 1005–1017.
30. Durbec, P.L., Larsson-Blomberg, L.B., Schuchardt, A., Costantini, F. and Pachnis, V. (1996) Common origin and developmental dependence on c-ret of subsets of enteric and sympathetic neuroblasts. *Development*, **122**, 349–358.
31. Golden, J.P., DeMaro, J.A., Osborne, P.A., Milbrandt, J. and Johnson, E.M., Jr (1999) Expression of neurturin, GDNF, and GDNF family-receptor mRNA in the developing and mature mouse. *Exp. Neurol.*, **158**, 504–528.
32. Le Douarin, N. and Kalcheim, C. (1999) *The Neural Crest*. Cambridge University Press, Cambridge, UK.
33. Nobrega, M.A., Ovcharenko, I., Afzal, V. and Rubin, E.M. (2003) Scanning human gene deserts for long-range enhancers. *Science*, **302**, 413.
34. Martin, N., Patel, S. and Segre, J.A. (2004) Long-range comparison of human and mouse Sprr loci to identify conserved noncoding sequences involved in coordinate regulation. *Genome Res.*, **14**, 2430–2438.
35. Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S. and Haussler, D. (2004) Ultraconserved elements in the human genome. *Science*, **304**, 1321–1325.
36. Jackson, I.J. and Abbott, C.M. (eds) (2000) *Mouse Genetics and Transgenics: a Practical Approach*. Oxford University Press, Oxford, New York.
37. Correia, K.M. and Conlon, R.A. (2001) Whole-mount *in situ* hybridization to mouse embryos. *Methods*, **23**, 335–338.