A multi-enhancer *RET* regulatory code is disrupted in Hirschsprung disease

Sumantra Chatterjee,[1,2,#+] Kameko M Karasaki,[2#] Ashish Kapoor,[2,3] and Aravinda Chakravarti [1,2,+]

[1] Center for Human Genetics and Genomics, New York University School of Medicine, New York, NY 10016; [2] Center for Complex Disease Genomics, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA; [3]Institute of Molecular Medicine, McGovern Medical School, University of Texas Health Science Center at Houston, Houston, TX 77030, USA.

# These two authors made equal contributions.

+Send all correspondence to:

Aravinda Chakravarti, Ph.D.
Center for Human Genetics & Genomics
New York University School of Medicine
435 East 30th Street, Room 802/3
New York, NY 10016
T: (212) 263-8023
E: aravinda.chakravarti@nyulangone.org

Sumantra Chatterjee, Ph.D.
Center for Human Genetics & Genomics
New York University School of Medicine
435 East 30th Street, Room 823R
New York, NY 10016
T: (212) 263-8029
E: sumantra.chatterjee@nyulangone.org

Abstract:

In Hirschsprung disease (HSCR; congenital colonic aganglionosis), three GWAS-identified common variants residing in three distinct enhancers of the *RET* receptor tyrosine kinase gene reduce its gene expression and are causal disease risk variants. Further, their combined effects significantly dysregulate the expression of other functionally-related genes defining the *RET* gene regulatory network (GRN). In this study, we asked how many variants in how many distinct *RET* enhancers affect HSCR risk by reducing *RET* gene expression? We demonstrate that 22 additional HSCR-associated polymorphisms, both independent and associated, reside within multiple candidate *RET* enhancers and among which 7 display differential allelic enhancer activities. Of these 7 *RET* enhancers, two bind PAX3 and extend the known *RET-EDNRB* GRN. Therefore, sequence variants within a minimum of 10 *RET* enhancers affect HSCR risk, revealing a diverse regulatory code modulating complex disease risk even at a single locus.

Introduction

Numerous genome wide association studies (GWAS) have convincingly demonstrated the role of noncoding variation in the primary etiology of complex traits and diseases. At any given locus, the associated variants are common, have individually small effects on the phenotypes and there are many associated polymorphisms making it unclear which of them are causal variants (Manolio et al., 2009). Even though complex traits and diseases are correctly understood to be comprised of the additive genetic effects across 100s to 1000s of loci distributed across the genome (Visscher et al., 2012), the detailed molecular genetic architecture at any individual component locus is unknown. Nevertheless, there are some general features of this architecture that we can deduce.

The causes of genetic associations at a locus, uncovered through GWAS, are common sequence variation within *cis* regulatory elements (CRE) or enhancers of a target gene modulating its gene expression and thereby contributing to the trait (Chakravarti and Turner, 2016). Thus, identification of all CREs within a locus would be immensely helpful in resolving the molecular basis of GWAS. This appears feasible since local gene expression control occurs within a topologically associating domain (TAD) (Dixon et al., 2012; Rao et al., 2014) limiting the physical locus within which such CREs can reside, and, there are many experimental methods to identify CREs comprehensively (Inoue et al., 2019; Kapoor et al., 2019). However, there are challenges as well because gene expression control varies spatio-temporally, by cell type and across development and aging, raising doubts whether all enhancers relevant to a phenotype can be identified (Chakravarti and Turner, 2016; Maurano et al., 2012) . Nevertheless, it is now possible to ask: how diverse is the collection of transcriptional enhancers with sequence variants at a locus contributing to a trait?

The above leads to an additional unanswered question: how can gene expression changes at one locus, presumably small, affect a trait comprising many such loci? One answer is that the different loci involved are functionally united through a gene regulatory network (GRN). Although the term GRN is variably used in genomics, the definition relevant here is the one by Eric Davidson (Davidson, 2010), i.e., a group of genes whose gene expression is regulated through transcriptional feedback, from genome to RNA and protein with feedback onto the genome. Thus, the expression changes in many independent genes get integrated through the GRN to accomplish a functional task. Our studies have demonstrated this GRN model for complex phenotypes (Chatterjee and Chakravarti, 2019; Chatterjee et al., 2016). Some authors contend that the vast majority, not all, of GWAS hits are irrelevant to the phenotype (Boyle et al., 2017) and hence the GRN hypothesis may help in elucidating the functional network.

In this study, we use Hirschsprung disease (HSCR; congenital colonic aganglionosis) to evaluate the diversity of *RET* enhancers and variants as causal risk factors, and their relationship to the known *RET-EDNRB* GRN. HSCR is a complex neurodevelopmental disorder in which failure of differentiation of enteric neural crest cell (ENCC) precursors in development leads to aganglionosis; more than 35 genes/loci explaining 67% of its population attributable risk have been implicated (Tilghman et al., 2019). Significantly, the majority of this risk arises from coding and enhancer variants of the *RET* receptor tyrosine kinase gene with minor contributions from other members of the *RET-EDNRB* GRN (Chatterjee and Chakravarti, 2019; Chatterjee et al., 2016; Kapoor et al., 2015; Tilghman et al., 2019). Consequently, here, we test all 30 HSCR-associated common (minor allele

frequency, MAF≥10%) variants, we previously identified in a GWAS (Jiang et al., 2015), to demonstrate that 22 of these lie within DNA elements with transcriptional enhancer activity and that 7 of these have alleles with differential enhancer activity in the neuroblastoma cell line SK-N-SH; 8 additional variants from this screen have been studied before (Chatterjee et al., 2016). We also demonstrate that two of these CREs bind the transcription factor (TF) PAX3, known to affect neural crest migration and enteric nervous system (ENS) differentiation (Bondurand et al., 2000; Lang et al., 2000); their knockdown significantly lowers *RET* transcription in SK-N-SH. Further, reducing *PAX3* gene expression reduces *SOX10* gene expression, the major TF in the *RET-EDNRB* GRN controlling *RET*. Therefore, *RET* is regulated in the ENS by at least 10 CREs with distinct causal variants for HSCR. The 10 CREs although functionally independent do not make independent contributions to HSCR owing to linkage disequilibrium. Nevertheless, multiple TF binding activities are disrupted to affect *RET* and *EDNRB* gene expression and HSCR risk through dysregulating the *RET-EDNRB* GRN. These results demonstrate how individually small enhancer allele effects contribute to an amplified response in conjunction with other enhancers of the same gene and the GRN to produce a genetic effect significant enough to disrupt development. We postulate this as a general mechanism for complex diseases.

Results

Enhancers at the RET locus:

To create a more complete catalog of the regulatory landscape of *RET* relevant to HSCR, we began with analysis of all 38 genome-wide significant non-coding common variants (minor allele frequency, MAF≥10%) discovered in a GWAS of 220 HSCR trios comprising a proband and both of her/his parents (Jiang et al., 2015). These 38 single

nucleotide polymorphisms (SNP) are distributed across a 155kb region (Chr10: 43434933-43590368; hg19) comprising the *RET* gene and arranged in 6 linkage disequilibrium (LD) blocks (**Figure 1A**) within the *RET* TAD; 8 of these elements have been studied before (Chatterjee et al., 2016). This region is epigenetically marked by 10 regions of DNase I hypersensitive (DHS), H3K4me1 and H3K27ac sites in 108-day human fetal intestine samples (Bernstein et al., 2010; Consortium, 2012) and 17 DHS regions in the neural crest-derived neuroblastoma cell line SK-N-SH (**Figure 1A**).

We conducted *in vitro* functional tests of enhancer activity of the remaining 30 SNPs by cloning ~500 bp elements centered on each SNP (**Table 1**) into a pGL4.23 luciferase vector with a minimal TATA-box of the β-globin gene and transfecting them into the SK-N-SH cell line. SK-N-SH expresses all known members of the *RET-EDNRB* GRN and is a cell model system for studies of the ENS transcriptional regulation (Chatterjee and Chakravarti, 2019; Chatterjee et al., 2016). Two pairs of SNPs (rs17158318/ rs17158320 and rs2506021/ rs2435342) were only 64bp and 108bp apart and hence were cloned into same elements (E9 and E24, respectively) (**Table 1**). As positive controls, we re-analyzed the 3 SNPs (rs2506030, rs7069590, rs2435357) known to demonstrate differential enhancer activity (Chatterjee et al., 2016). The results showed that 22 elements had significant enhancer activity (P <0.001 and >2X reporter activity over the promoter only control vector) but only 7 elements (E2, E4, E5, E14, E26, E27 and E28) also had differential allelic reporter activity (**Table 1 and Figure 1B**). Of these, 5 or 71% (E2, E4, E26, E27 and E28) overlapped an open chromatin region or an enhancer-associated epigenetic mark, as compared to 3 (E9, E16, E21) out of 15 (20%) elements among the remainder which showed reporter activity

but no allelic difference (**Figure 1A**). Thus, E2, E4, E26, E27 and E28 have compelling while E5 and E14 have strong evidence of being enhancers in the human fetal gut, i.e., at least 10 functionally distinct CREs affect HSCR.

Haplotype specific effect of causal RET polymorphisms:

The identification of 10 (rs788263, rs788261, rs788260, rs2506030, rs1547930, rs7069590, rs2435357, rs12247456, rs7393733 and rs2505541) CRE-associated risk variants allowed us to ask which allelic combinations were associated with disease risk. First, we estimated haplotypes and their frequencies for all 10 variants in 220 unrelated HSCR cases of European-ancestry (Jiang et al., 2015) and 503 unrelated controls of non-Finnish European-ancestry from the 1000 Genomes project (Genomes Project et al., 2015). Second, we calculated the odds ratio (OR) for all haplotypes with a frequency of ≥1% in controls. We observed 10 distinct haplotypes of which CTGAACCAC**T** (risk allele in bold) was used as the reference because it had the smallest number (one) of risk alleles (**Table 2**). Significant odds ratios (OR) were observed for 2 haplotypes relative to the reference: significant increases for **GCAGGTTGGT** (OR 12.2, 95% CI: 5.97-24.93, $P=7.02x10^{-12}$) and CTGA**GTTGGT** (OR 7.2, 95% CI: 3.26-15.91, $P=1.02x10^{-6}$) (**Table 2**). Note that these two haplotypes contained our previously identified risk increasing ATT and GTT haplotypes (for rs2506030, rs7069590, rs2435357) (Chatterjee et al., 2016). Our data precludes accurate risk estimation for all haplotypes but, in general, HSCR risk increases with increasing numbers of risk alleles. These results suggest that risk of or protection from HSCR depends on the simultaneous binding or lack-of-binding of multiple independent TFs at *RET* CREs, thereby revealing a *RET* regulatory code.

Transcription factors regulating RET:

Enhancers exerts their control on gene expression via binding of specific TFs whose identities are crucial for deciphering the GRN. We searched for TF binding sites (TFBS) within the 7 novel risk-associated CREs using publicly available databases and software, and a stringent cut off for TF discovery ($P< 10^{-4}$), and identified four candidate TFs: (1) PAX3 binding to E2 (AATAAACCC; $P=4.67x10^{-5}$); (2) ZBTB6 to E5 (TGGCTCCATCATG; $P=2.387x10^{-6}$); (3) ZNF263 binding to E14 (GCCTCACTGCTCCAG; $P=8.09x10^{-5}$) and (4) PAX3 at E27 (TCGTCACTCTTAC; $P= 9.99x10^{-5}$). To determine their relevance for HSCR, we performed qPCR in SK-N-SH cells and observed no expression for *ZNF263* and *ZBTB6* (data not shown). Further, their expression was absent in the developing mouse gut (Chatterjee et al., 2019), where *RET* expression is critical for ENS development, making it highly unlikely that these TFs control *RET* expression via specific enhancers. Moreover, *ZNF263* and *ZBTB6* are both zinc finger domain-containing proteins that are GC-rich and overestimate their statistical significance owing to their rarity.

In contrast, we detected robust expression of *PAX3 i*n both SK-N-SH and the developing mouse gut (Chatterjee et al., 2019). Thus, we performed ChIP-qPCR for *PAX3* in SK-N-SH cells and detected significant binding at both E2 (18-fold enrichment; $P= 10^{-3}$) and at E27 (26-fold enrichment, ($P= 5x10^{-4}$) (**Figure 2**). We further demonstrated the specificity of this binding by performing ChIP-qPCR after siRNA-mediated knockdown of PAX3 in SK-N-SH cells to show a 1.3-fold ($P=8x10^{-3}$) reduced binding at E2 and 2-fold ($P=4x10^{-4}$) reduced binding at E27 (**Figure 2**). To prove that PAX3 does indeed control *RET* gene expression, we also measured *RET* levels after siRNA-mediated knockdown of *PAX3*. As positive controls, we also measured

RET levels after siRNA-mediated knockdown of other established *RET* TFs, namely SOX10, GATA2 and RARB, to recapitulate known effects in the *RET-EDNRB* GRN.

These experiments showed that decreasing *PAX3* lead to a 49% ($p=4x10^{-4}$) reduction in *RET* in comparison to 76% ($p=2.3x10^{-6}$), 50% ($p= 3.1x10^{-3}$) and 81% ($p= 4.1x10^{-5}$) decreases consequent to *SOX10*, *GATA2* and *RARB* knockdown, respectively; as a control, knockdown of *RET* by its specific siRNA reduced its expression by 96% ($p=4.4x10^{-6}$) (**Figure 3A**). We had previously demonstrated that there is considerable cross talk between the various RET TFs (Chatterjee et al., 2016). Consequently, we measured gene expression of *SOX10*, *GATA2* and *RARB* after siRNA-mediated knockdown of *PAX3*: we observed only a significant drop in *SOX10* gene expression (32% decrease, $P= 3x10^{-3}$) but both *GATA2* and *RARB* levels were unchanged (**Figure 3B)**.

in vitro evidence of RET enhancers:

The human genetic evidence for HSCR-associated polymorphisms within transcriptional enhancers identified from *in vitro* (reporter activity) and *ex vivo* (siRNA in SK-N-SH) experiments could be buttressed by additional data. The most direct would be by deletion analysis of each enhancer rather than knockdown, although knockdown but not deletion of multiple CREs is feasible. To demonstrate this benefit, we deleted each individual enhancer identified in our initial screen, RET-7, RET-5.5 and RET+3 which are bound by RARB, GATA2 and SOX10, respectively (Chatterjee et al., 2016). We used paired guide RNAs upstream and downstream of the 3 target enhancers to direct espCAS9 to the specific regions to introduce deletions and non-homologous end joining in SK-N-SH cells (**Figure 4**). We screened 48 independently transfected wells for each enhancer region and detected only heterozygous

deletions for each. For RET-7, we identified 2 clones (4%) with 169bp and 183bp deletions, both of which deleted the SNP rs2506030. We detected 3 heterozygous clones for RET-5.5 deletion (6.25%) with 2 clones with 147bp and 1 clone with a 139bp, all of which deleted the SNP rs7069590. Similarly, we also detected 3 clones (6.25%) for RET+3 deletion with 1 clone encompassing 147bp and 2 clones encompassing 138bp deletions, all of which deleted the SNP rs2435357. (**Figure 4**).

For analysis, we measured gene expression for *RET*, *EDNRB*, *SOX10*, *GATA2* and *RARB* by qPCR in these clones to detect the effect of heterozygous enhancer deletions on the RET GRN. We observed that deletion of RET+3 lead to 32% ($P=3.1x10^{-3}$) loss of *RET* expression compared to cells transfected with only a Cas9 vector without the guide RNA; the corresponding decreases for RET-5.5 and RET-7 are 22% ($P=2.8X10^{-3}$) and 20% ($P=2.6X10^{-3}$), respectively (**Figure 4**). Thus, at least in SK-N-SH these 3 CREs can act as direct enhancers of *RET* transcription. Further, deletion of RET+3 has a more severe effect on *RET* expression as compared to the other two, as also seen from their disease risk (Chatterjee et al., 2016; Emison et al., 2010; Kapoor et al., 2015). However, in these experiments, the remaining genes of the GRN are unaffected transcriptionally, which is not unexpected given our previous data that *RET* gene expression has to fall below 50% before any GRN transcriptional effect is observed (Chatterjee and Chakravarti, 2019; Chatterjee et al., 2016)

Discussion

It is a truism that understanding the function of a gene is critical for understanding its role in normal physiology and how its disruption leads to disease. Although for decades human

disease studies, primarily of Mendelian disorders, have focused on analysis of genes and their products directly, today, for complex traits and disorders, it is imperative to understand the regulatory architecture of a gene. Two critical components of this architecture are identifying a gene's CREs and the TFs that binds to them. The studies reported here, for Hirschsprung disease, indicate that even for a disorder specific to particular cells, such as ENCCs, regulation of a critical gene such as *RET* involves at least 10 enhancers. Interestingly, the new TF identified, PAX3, is also a component of the known *RET-EDNRB* GRN.

These studies demonstrate the extensive regulatory control of *RET*. Although all 10 enhancers affect *RET* gene expression, and some of them may be shadow enhancers with the same function, others may act at different developmental stages, as we have demonstrated before for RARB, GATA2 and SOX10 enhancers (Chatterjee et al., 2016), or act on different groups of ENCCs with different functions (e.g., proliferation potential). Thus, we need broader enhancer screens to define the full enhancer architecture of *RET*; perhaps this will reveal yet more CREs. We also need to elucidate the full repertoire of TFs that regulate *RET*, requiring experimental analysis beyond bioinformatic searches. These pieces of information are crucial to understand the full composition of the *RET-EDNRB* GRN and, therefore, the mutational targets of HSCR.

The human genetics implications of these data, beyond understanding HSCR, are that genetically independent (unassociated) SNPs at a specific GWAS locus are not the only candidate variants for understanding the phenotype. Additional SNPs are surely involved, even those perfectly associated with one another provided they affect functionally independent

enhancers. This suggests that understanding the regulatory contributions to GWAS will require experimental data on enhancer effects in a phenotype beyond what statistical analysis can provide.

HSCR is a “RETopathy” because the vast majority of its risk arises from the *RET-EDNRB* GRN and in which *RET* can be dysregulated by variants within its coding sequence, its enhancers or other genes that interact with it directly. However, this GRN needs to account for the many TFs that have a critical role in ENS formation and with transcriptional effects on and from *RET* (Chatterjee et al., 2019). Chief among them is PAX3, where multiple lines of evidence emphasizes its critical role in neural crest cell migration and differentiation and in enteric neurons (Bondurand et al., 2000; Lang et al., 2000). Now we show that PAX3 regulates RET directly through two independent enhancers. Previous studies have also shown interactions between PAX3 and SOX10 (Lang and Epstein, 2003) which we also identify here.

Given RET’s regulatory diversity why is HSCR risk so variable across individuals and why does it scale with increasing numbers of risk alleles. Based on our results we suggest that this arises due to *RET* gene expression decreasing proportional to the enhancer dosage defect. Proving this hypothesis will require creating cellular and animal models with differing numbers of enhancer mutations. This is preferable to creating multiple TF mutations to control cell-type specific developmental defects. We have embarked on constructing such models using methods from synthetic biology (Richardson et al., 2017).

Materials and methods

ChIP-seq peak calling:

Three epigenomic datasets from a 108-day human fetal large intestine, histone H3K27ac ChIP-seq (GSM1058765), histone H3K4me1 ChIP-seq (GSM1058775), and DNaseI-seq (GSM817188), were downloaded from the NIH Roadmap Epigenomics Project. For the SK-N-SH cell line, DNaseI-seq data (GSM736559) were obtained from the ENCODE project. For each dataset, MACS software v1.4 (Zhang et al., 2008) with default settings was used to call ''peaks'' where sequence reads were significantly enriched over background. With the default peak-calling threshold ($p < 10^{-5}$), 51,771, 61,689, 66,930 and 52,534 genomic regions were identified in the GSM1058765, GSM1058775, GSM817188 and GSM736559 datasets, respectively.

Reporter assays:

400 ng of firefly luciferase vector (pGL4.23, Promega Corporation,USA) containing the DNA sequence of interest and 2 ng of Renilla luciferase vector (transfection control) were transiently transfected into cell line SK-N-SH (5 – $6x10^4$ cells/well), using 6 ml of FuGENE HD transfection reagent (Roche Diagnostic, USA) in 100ml of OPTIMEM medium (Invitrogen, USA). Cells were grown for 48 hours and luminescence measured using a Dual Luciferase Reporter Assay System on a Tecan multi-detection system luminometer, per the manufacturer's instructions.

Cell lines:

The human neuroblastoma cell SK-N-SH, purchased from ATCC, USA (# HTB-11), was grown under standard conditions (DMEM + 10% FBS and 1% Penicillin Streptomycin). It was maintained in 10 cm culture dishes and passaged every 48 hours when it reached ~80% confluency.

Chromatin Immunoprecipitation–qPCR (ChIP-qPCR) assays:

ChIP was performed thrice independently for each antibody using $1x10^6$ SK-N-SH cells for each transcription factor using the EZ-Magna ChIP kit (Millipore Sigma, USA), as per the manufacturer's instruction, with the following modifications: the chromatin was sonicated for

30s on and 30s off for 10 cycles; sheared chromatin was pre-blocked with unconjugated beads for 4 hours and specific antibodies separately conjugated to the beads for 4 hours before immunoprecipitation was performed with the pre-blocked chromatin. A polyclonal antibody was used against PAX3 (16HCLC; Invitrogen, USA) at 15μg concentration. ChIP assays were also performed on cells 48 hr after transfection with the PAX3 siRNAs (L-012399-00-0005; Dharmacon/Horizon Discovery, USA) at 25 μM to assess the specificity of TF binding. qPCR assays were performed using SYBR green (Life Technologies, USA) and specific primers against enhancer E2 (E2_FWD 5' GCTGCAGATATGCAACTTCCAA 3' and E2_REV 5' AGATATGCTGGTGAGGGGCT 3') and enhancer E27 (E27_FWD 5' AGGAAGGTAGGCACCCTGTA and E27_REV 5' AGCCCTGTGTTAACTGTCCG 3'). The data were normalized to input DNA and enrichment was calculated by fold excess over ChIP performed with specific IgG as background signal. All assays were done in triplicate for each independent ChIP assay (n = 9).

siRNA assays:

PAX3 (L-012399-00-0005), *RET* (L-003170-00-0005), *SOX10* (L-017192-00), *GATA2* (L-009024-02 and *RARB* (L-003438-02) SMARTpool siRNAs (a combination of 4 individual siRNAs targeting each gene) along with ON-TARGET plus non-targeting siRNAs (D-001810-10, negative control) (Dharmacon/ Horizon Discovery, USA) were transfected at 20 μM in SK-N-SH cells at a density of 10^4 -10^5 cells using FuGene HD Transfection reagent (Promega Corporation, USA) per the manufacturer's instructions. Negative control siRNAs were always transfected at 25 μM concentration.

Gene expression assays:

Total RNA was extracted from SK-N-SH cells using TRIzol (Life Technologies, USA) and cleaned on RNeasy columns (QIAGEN, USA). 500μg of total RNA was converted to cDNA using SuperScriptIII reverse transcriptase (Life Technologies, USA) using Oligo-dT primers. The diluted (1/5) total cDNA was subjected to Taqman gene expression (ThermoFisher Scientific, USA) using the following transcript-specific probes and primers: *RET* (Hs01120032_m1), *EDNRB* (Hs00240747_m1), *PAX3*(Hs00240950_m1), *SOX10* (Hs00366918_m1), *GATA2* (Hs00231119_m1) and *RARB* (Hs00977140_m1)

Human β-actin was used as an internal loading control, as appropriate, for normalization. For siRNA knockdown experiments, five independent wells for SK-N-SH cells were used for RNA extraction and each assay performed in triplicate (n = 15). Relative fold change was calculated based on the 2DDCt (threshold cycle) method. For siRNA experiments, 2DDCt for negative control non- targeting control siRNA was set to unity. P values were calculated from pairwise 2-tailed t tests and the data presented as means with their standard errors (SE). Subsequently, P-values were calculated from pairwise 2-tailed t tests data is presented as the mean fold change with its standard error.

CRIPSR/Cas9 induced deletions:

The following genomic regions were targeted with pairs of guides for 3 enhancers: RET-7 (*RET-7_DGR- 5'GACCCCTCTGGACCCGTCGCGG3 and RET-7_UGR- CCGAGTGTTTGCGCCCATGAGG*), RET-5.5 (*RET-5.5_DGR- AAGGATGGATACACCTTCCGGG and RET-5.5_UGR- AAGGATGGATACACCTTCCGGG*) and RET+3 (*RET+3_DGR- GCCTGGCTGGAGGTCCAAGAGG and RET+3_UGR- GCTCATGAGGAGCACACCGTGG*). The guide RNA and CAS9 were cloned into a single vector ***U6-gRNA/CMV-Cas9-GFP*** (Millipore Sigma, USA) and each pair of guides were transfected into 72 wells of low-density (~10,000 cells) plated SK-N-SH cells. We also transfected 72 wells with only the Cas9-GFP vector without the guide RNA to serve as transcriptional control against which to measure change in gene expression due to enhancer deletions.

After 24 hours cells were checked for GFP expression and wells with at least ~80% cells expressing GFP were re-plated, ensuring high transfection efficiency. For all 3 constructs and the control we obtained at least 48 wells with over 80% transfection efficiency. The cells were further grown for 24 hours and then split into 2 tubes of equal cells. One was used for DNA extraction by standard phenol-chloroform methods and other used for RNA extraction.

To check for deletions, we used the following pairs of primers: RET-7 (*RET-7_FWD 5' GAGGAACCGCGCACATCAG 3' and RET-7_REV 5' CGCCTTGCCTGGCCCTCCA 3'*), RET-5.5 (*RET-5.5_FWD 5' GTCCTGGTTGCCCCACTCTG 3' and RET-5.5_REV 5' GGTCAGGTGTCACAAAGTCT 3'*) and RET+3 (*RET+3_FWD 5'*

AATGGGCAAGACCATCTCAG 3' and RET+3_*REV 5' GTGGCCAGTAGCTGGAAGAGCAG 3*'). The primers amplify a wildtype sequence for 1.1kb for each enhancer region. Each cell line was PCR confirmed for enhancer deletion and Sanger sequenced to re-confirm the exact region of deletion. For quantifying gene expression changes RNA was extracted from each cell line with deletion and qPCR assays performed in triplicate; n=6 for RET-7 and n=9 for RET-5.5 and RET+3. Quantitative gene expression changes were calculated as detailed above.

Estimating haplotype-specific HSCR risk:

Genotypes at the 10 RET functional variants in 220 S-HSCR cases and 503 European-ancestry controls were obtained from the our published Genome Wide Association study (Jiang et al., 2015) and the 1000 Genomes project (Genomes Project et al., 2015), respectively. Haplotypes were generated from unphased genotype data using BEAGLE (Browning and Browning, 2007) and filtered to retain only those that had a frequency >1% in cases and controls. Standard methods using $\chi 2$ statistics were used for calculation of haplotype-count based odds ratios (OR), their upper and lower confidence limits and significance of their deviation from the null hypothesis of no association (OR = 1) (Kapoor et al., 2015).

Identifying transcription factors for candidate enhancers:

We searched for TF binding sites (TFBS) within risk-associated CREs using FIMO (Bailey et al., 2009; Grant et al., 2011) and 890 validated TF motifs in TRANSFAC (Wingender et al., 1996). We used the setting of "minimize false positives" and a stringent cut off of $P< 10^{-4}$ to detect putative TF binding.

Author Contributions:

S.C and A.C. conceived and designed the study. K.M.K. conducted all *in vitro* luciferase assays, A.K. all genotyping assays and S.C all in vivo CRISPR assays. S.C. and A.C. wrote the manuscript.

Funding:

This work was supported by a National Institute of Health MERIT Award HD28088 to A.C.

REFERENCES

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res *37*, W202-208.

Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M.A., Beaudet, A.L., Ecker, J.R.*, et al.* (2010). The NIH Roadmap Epigenomics Mapping Consortium. Nat Biotechnol *28*, 1045-1048.

Bondurand, N., Pingault, V., Goerich, D.E., Lemort, N., Sock, E., Le Caignec, C., Wegner, M., and Goossens, M. (2000). Interaction among SOX10, PAX3 and MITF, three genes altered in Waardenburg syndrome. Hum Mol Genet *9*, 1907-1917.

Boyle, E.A., Li, Y.I., and Pritchard, J.K. (2017). An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell *169*, 1177-1186.

Browning, S.R., and Browning, B.L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet *81*, 1084-1097.

Chakravarti, A., and Turner, T.N. (2016). Revealing rate-limiting steps in complex disease biology: The crucial importance of studying rare, extreme-phenotype families. Bioessays *38*, 578-586.

Chatterjee, S., and Chakravarti, A. (2019). A gene regulatory network explains RET-EDNRB epistasis in Hirschsprung disease. Hum Mol Genet *28*, 3137-3147.

Chatterjee, S., Kapoor, A., Akiyama, J.A., Auer, D.R., Lee, D., Gabriel, S., Berrios, C., Pennacchio, L.A., and Chakravarti, A. (2016). Enhancer Variants Synergistically Drive Dysfunction of a Gene Regulatory Network In Hirschsprung Disease. Cell *167*, 355-368 e310.

Chatterjee, S., Nandakumar, P., Auer, D.R., Gabriel, S.B., and Chakravarti, A. (2019). Gene- and tissue-level interactions in normal gastrointestinal development and Hirschsprung disease. Proc Natl Acad Sci U S A.

Consortium, E.P. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57-74.

Davidson, E.H. (2010). Emerging properties of animal gene regulatory networks. Nature *468*, 911-920.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376-380.

Emison, E.S., Garcia-Barcelo, M., Grice, E.A., Lantieri, F., Amiel, J., Burzynski, G., Fernandez, R.M., Hao, L., Kashuk, C., West, K.*, et al.* (2010). Differential contributions of rare and common, coding and noncoding Ret mutations to multifactorial Hirschsprung disease liability. Am J Hum Genet *87*, 60-74.

Genomes Project, C., Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A.*, et al.* (2015). A global reference for human genetic variation. Nature *526*, 68-74.

Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: scanning for occurrences of a given motif. Bioinformatics *27*, 1017-1018.

Inoue, F., Kreimer, A., Ashuach, T., Ahituv, N., and Yosef, N. (2019). Identification and Massively Parallel Characterization of Regulatory Elements Driving Neural Induction. Cell Stem Cell *25*, 713-727 e710.

Jiang, Q., Arnold, S., Heanue, T., Kilambi, K.P., Doan, B., Kapoor, A., Ling, A.Y., Sosa, M.X., Guy, M., Jiang, Q.*, et al.* (2015). Functional loss of semaphorin 3C and/or semaphorin 3D and their epistatic interaction with ret are critical to Hirschsprung disease liability. Am J Hum Genet *96*, 581-596.

Kapoor, A., Jiang, Q., Chatterjee, S., Chakraborty, P., Sosa, M.X., Berrios, C., and Chakravarti, A. (2015). Population variation in total genetic risk of Hirschsprung disease from common RET, SEMA3 and NRG1 susceptibility polymorphisms. Hum Mol Genet *24*, 2997-3003.

Kapoor, A., Lee, D., Zhu, L., Soliman, E.Z., Grove, M.L., Boerwinkle, E., Arking, D.E., and Chakravarti, A. (2019). Multiple SCN5A variant enhancers modulate its cardiac gene expression and the QT interval. Proc Natl Acad Sci U S A *116*, 10636-10645.
Lang, D., Chen, F., Milewski, R., Li, J., Lu, M.M., and Epstein, J.A. (2000). Pax3 is required for enteric ganglia formation and functions with Sox10 to modulate expression of c-ret. J Clin Invest *106*, 963-971.

Lang, D., and Epstein, J.A. (2003). Sox10 and Pax3 physically interact to mediate activation of a conserved c-RET enhancer. Hum Mol Genet *12*, 937-945.
Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A.*, et al.* (2009). Finding the missing heritability of complex diseases. Nature *461*, 747-753.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J.*, et al.* (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science *337*, 1190-1195.
Rao, S.S., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S.*, et al.* (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell *159*, 1665-1680.

Richardson, S.M., Mitchell, L.A., Stracquadanio, G., Yang, K., Dymond, J.S., DiCarlo, J.E., Lee, D., Huang, C.L., Chandrasegaran, S., Cai, Y.*, et al.* (2017). Design of a synthetic yeast genome. Science *355*, 1040-1044.

Tilghman, J.M., Ling, A.Y., Turner, T.N., Sosa, M.X., Krumm, N., Chatterjee, S., Kapoor, A., Coe, B.P., Nguyen, K.H., Gupta, N.*, et al.* (2019). Molecular Genetic Anatomy and Risk Profile of Hirschsprung's Disease. N Engl J Med *380*, 1421-1432.
Visscher, P.M., Brown, M.A., McCarthy, M.I., and Yang, J. (2012). Five years of GWAS discovery. Am J Hum Genet *90*, 7-24.

Wingender, E., Dietze, P., Karas, H., and Knuppel, R. (1996). TRANSFAC: a database on transcription factors and their DNA binding sites. Nucleic Acids Res *24*, 238-241.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol 9, R137

TABLES

Genomic element	SNP ID	Genomic location (chr10)		Non-Risk Allele	Risk Allele
		Start	End		
E1	rs788267	43434932	43435936	C	T
E2	**rs788263**	43437007	43437506	C	G
E3	rs788262	43437440	43437943	A	G
E4	**rs788261**	43437726	43438225	T	C
E5	**rs788260**	43438228	43438727	G	A
E6	rs2995411	43440554	43441053	C	T
E7	rs1582227	43441446	43441952	T	C
E8	rs2488278	43446082	43446581	T	C
E9	rs17158318/ rs17158320	43448581	43449090	G/C	A/A
E10	rs947696	43455083	43455582	G	T
E11	rs7908085	43460567	43461065	T	A
E12	rs10900290	43471322	43471823	C	T
E13	rs947690	43479479	43479979	G	C
E14	**rs1547930**	43483056	43483559	G	A
E15	rs3004258	43483900	43484403	T	G
E16	rs3026703	43557546	43558048	T	C
E17	rs3026707	43558368	43559122	A	G
E18	rs2505989	43563195	43563700	G	C
E19	rs741763	43568087	43568586	G	C
E20	rs2505995	43569379	43569878	A	G
E21	rs2506010	43573167	43574025	C	T
E22	rs2506020	43578754	43579373	C	T
E23	rs2506021/ rs2435342	43583869	43584408	C/T	T/C
E24	rs2506022	43584264	43584845	C	T
E25	rs2435343	43585384	43585889	T	G
E26	**rs12247456**	43587982	43588420	G	A
E27	**rs7393733**	43588440	43588915	G	C
E28	**rs2505541**	43589862	43590368	T	C

Table 1: Genomic coordinates (hg19) of 28 elements containing 30 single nucleotide polymorphisms (SNPs) at the *RET* locus which are associated with HSCR. Elements in bold act as enhancers in *in vitro* transcriptional assays while SNPs in bold show allelic difference between the risk and non-risk allele as well.

Haplotype	Case Frequency	Control Frequency	Odds Ratio (95% CI)	P
CTGAACCACT	0.02	0.08	1	1
CTGAATCACT	0.04	0.07	2.24 (0.95-5.29)	0.07
CTGAGCCACT	0.03	0.10	1.13 (0.45-2.82)	0.79
CTGAATCGGC	0.04	0.06	2.84 (1.20-6.77)	0.02
CTGAGTCGGC	0.06	0.16	1.58 (0.71-3.51)	0.26
CTGAATTGGT	0.02	0.01	4.35 (1.41-13.49)	0.01
CTGAGTTGGT	0.10	0.06	**7.2 (3.26-15.91)**	**1.02×10^{-6}**
GCAGGTCACT	0.02	0.04	2.45 (0.92-6.54)	0.07
GCAGGTCGGC	0.06	0.15	1.62 (0.72-3.61)	0.24
GCAGGTTGGT	0.52	0.17	**12.20 (5.97-24.93)**	**7.02×10^{-12}**

Table 2: Haplotypes for ten *RET* enhancer polymorphisms (rs788263, rs788261, rs788260, rs2506030, rs1547930, rs7069590, rs2435357, rs12247456, rs7393733 and rs2505541), with green and red denoting the non-risk and risk allele, respectively. Observed frequencies of haplotypes among 220 European ancestry HSCR cases and 503 controls together with the odds ratio (with respect to the reference haplotype CTGAACCACT containing only one susceptibility allele; significant values bolded) and statistical significance (P) are shown.

FIGURES

Figure 1: **The *RET* regulatory landscape in the enteric nervous system.** (A) The155kb *RET* locus (10: 43434933- 43590368; hg19) contains 38 putative cis regulatory elements (CRE) with HSCR-associated polymorphisms in 4 linkage disequilibrium (LD) blocks. Multiple enhancer-associated epigenetic marks (DHS: DNaseI hypersensitivity, H3K27ac, H3K4me1) were detected in 108-day human fetal large intestine and the SK-N-SH neuroblastoma cell line, and transcription factor (TF) binding sites. Common (≥10% allele frequency) variants associated with HSCR as well as showing allelic difference in *in vitro* transcription assays are marked in red. (B) *in vitro* luciferase assays of 30 elements in (A) in SK-N-SH cells. 22 putative CREs act as enhancers as compared to a promoter-only control, of which 7 contained a HSCR-associated polymorphism with allelic difference in luciferase activity (boxed). We also re-analyzed the three positive controls RET-7, RET-5.5 and RET+3. Error bars are standard errors (SE) of the mean (*P < 0.01, **P < 0.001) for 3 independent biological replicates for each assay.

Figure 2: Identification of the cognate transcription factors bound to *RET* enhancers.
Genomic map of the *RET* locus with locations of CREs E2 and E27 together with ChIP-qPCR results using a PAX3 antibody showing enrichment of binding as compared to the background in SK-N-SH cells. Specificity of this binding is shown by siRNA knockdown of *PAX3* with a concomitant reduction in the ChIP-qPCR signal at both CREs. Error bars are standard errors (SE) of the mean (*P < 0.01, **P < 0.001) for 3 independent biological replicates for each ChIP assay.

Figure 3: TF-mediated *in vitro* and *in vivo* control of gene expression.
(A) siRNA-mediated knockdown of *PAX3*, *SOX10*, *GATA2*, *RARB and RET* in SK-N-SH cells downregulates *RET* transcription. (B) siRNA-mediated knockdown of PAX3 has significant transcriptional effects on *SOX10* but not on *GATA2* or *RARB*. Error bars are standard errors (SE) of the mean (*P < 0.01, **P < 0.001) for 5 independent biological replicates in all experiments.

Figure 4: Deletion of *RET* enhancers affects *RET* transcription *in vitro*.
A genomic map of the *RET* locus showing 3 CREs and their corresponding HSCR associated polymorphisms: RET-7 (rs2506030), RET-5.5 (rs7069590) and RET+3 (rs2435357). The green boxed region represents the deleted region in each of the enhancers as confirmed by Sanger sequencing and validated on an agarose gel. Heterozygous deletion of all 3 enhancers leads to significant loss of *RET* expression but does not affect the other genes of the GRN. Error bars are standard errors (SE) of the mean (*P < 0.01, **P < 0.001) across 6 replicates for RET-7 and 9 replicates for RET-5.5 and RET+3. UGR: upstream guide RNA, DGR: downstream guide RNA

FIGURE 1

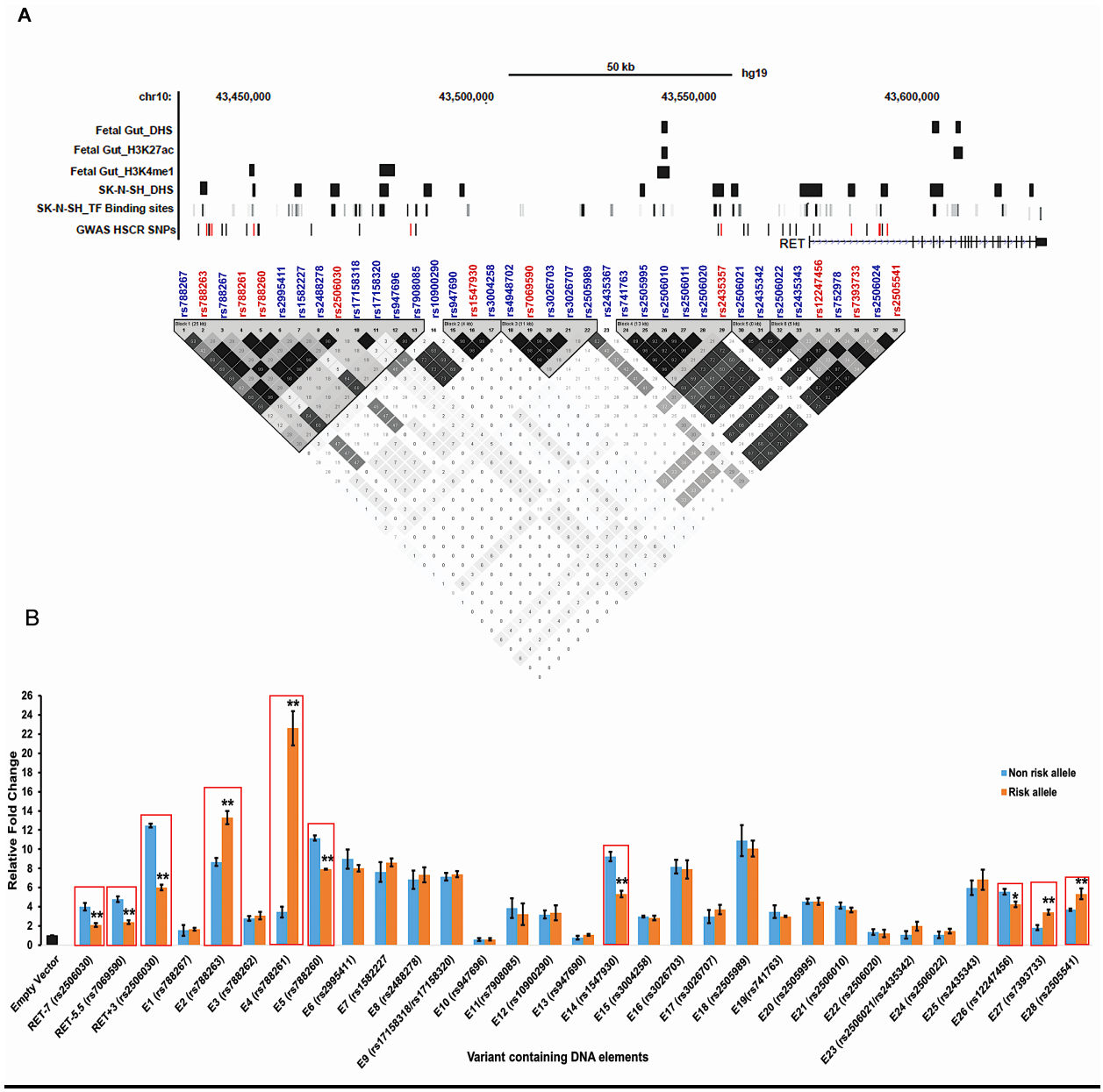

A
50 kb
hg19
chr10:
43,450,000
43,500,000
43,550,000
43,600,000
Fetal Gut_DHS
Fetal Gut_H3K27ac
Fetal Gut_H3K4me1
SK-N-SH_DHS
SK-N-SH_TF Binding sites
GWAS HSCR SNPs
RET
B
Relative Fold Change
Non risk allele
Risk allele
Empty Vector
RET-7 (rs2506030)
RET-5.5 (rs7069590)
RET+3 (rs2506030)
E1 (rs788267)
E2 (rs788263)
E3 (rs788262)
E4 (rs788261)
E5 (rs788260)
E6 (rs2995411)
E7 (rs1582227)
E8 (rs2488278)
E9 (rs17158318/rs17158320)
E10 (rs947696)
E11(rs7908085)
E12 (rs10900290)
E13 (rs947690)
E14 (rs1547930)
E15 (rs3004258)
E16 (rs3026703)
E17 (rs3026707)
E18 (rs2505989)
E19(rs741763)
E20 (rs2505995)
E21 (rs2506010)
E22 (rs2506020)
E23 (rs2506021/rs2435342)
E24 (rs2506022)
E25 (rs2435343)
E26 (rs12247456)
E27 (rs7393733)
E28 (rs2505541)
Variant containing DNA elements

FIGURE 2

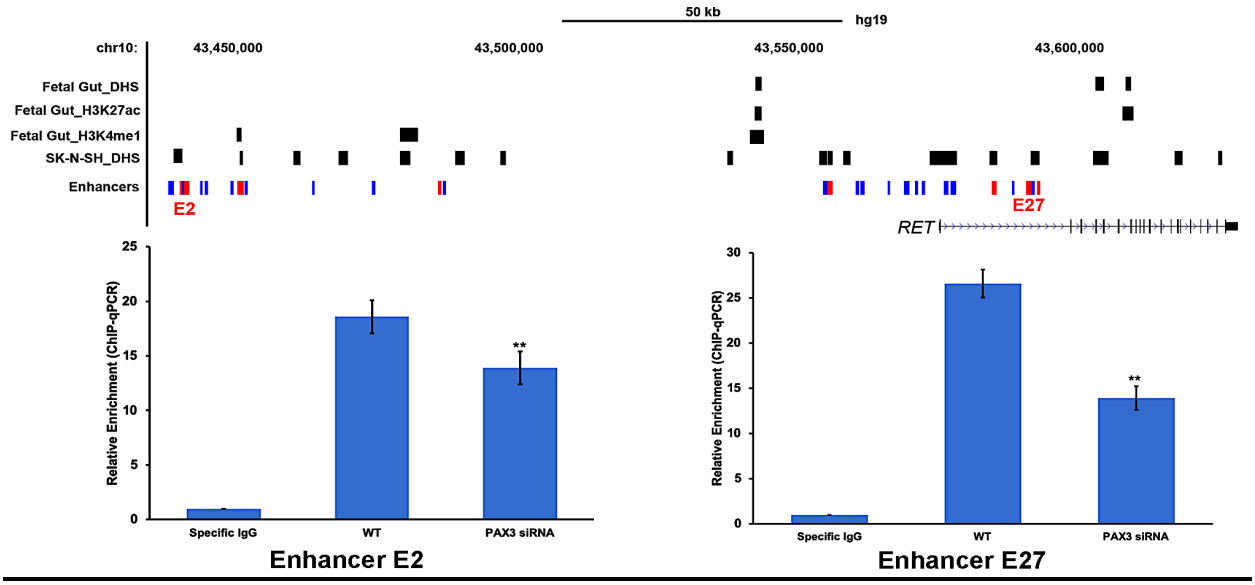


FIGURE 3

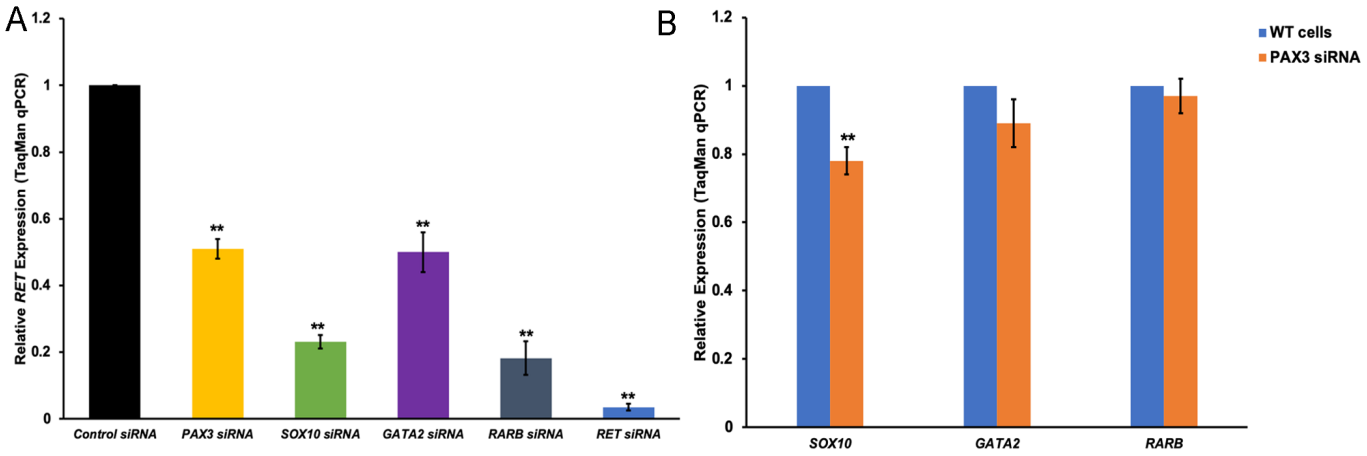
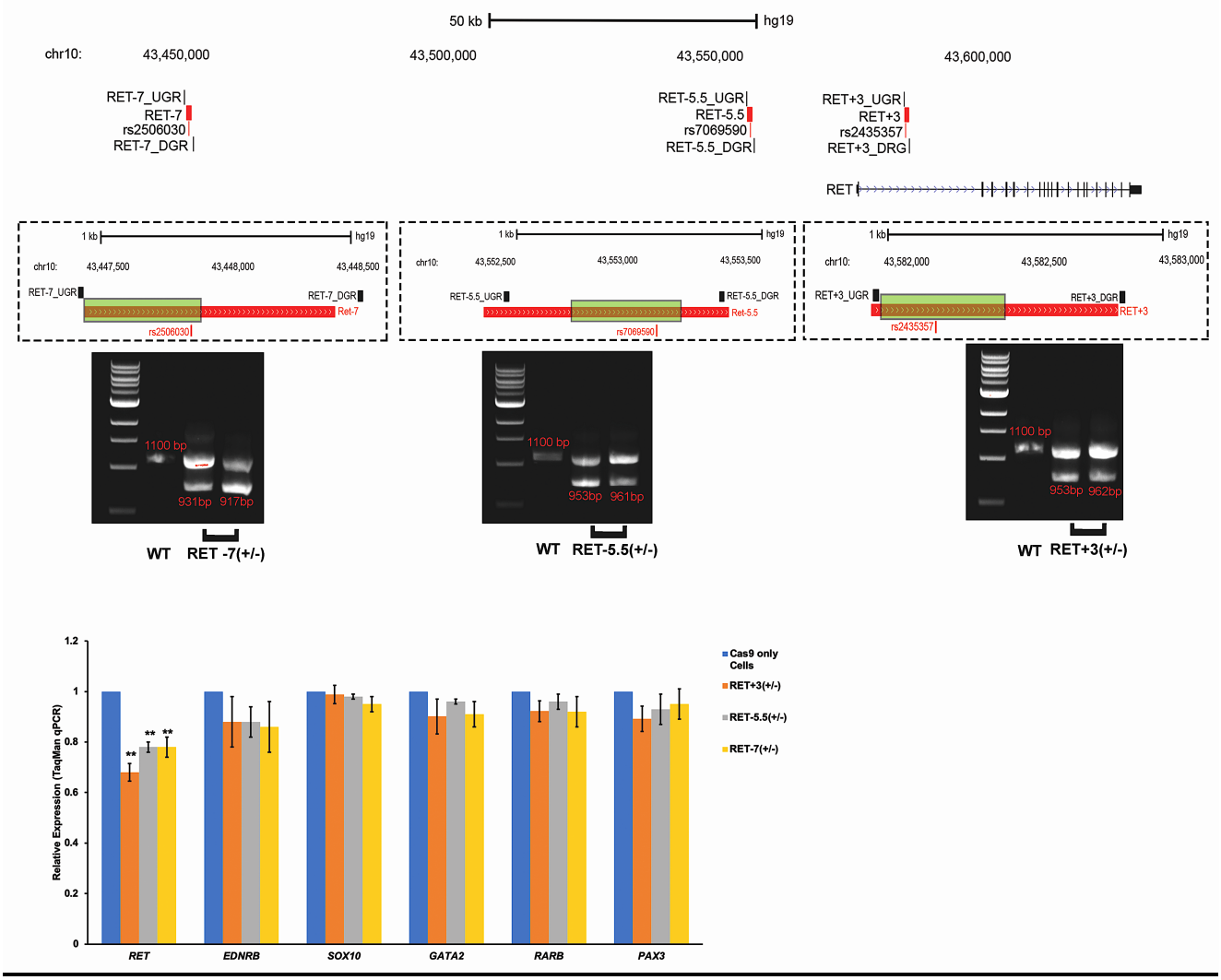


FIGURE 4


50 kb
hg19
chr10:
43,450,000
43,500,000
43,550,000
43,600,000
RET-7_UGR
RET-7
rs2506030
RET-7_DGR
RET-5.5_UGR
RET-5.5
rs7069590
RET-5.5_DGR
RET+3_UGR
RET+3
rs2435357
RET+3_DRG
RET
1 kb
hg19
chr10:
43,447,500
43,448,000
43,448,500
RET-7_UGR
RET-7_DGR
Ret-7
rs2506030
1 kb
hg19
chr10:
43,552,500
43,553,000
43,553,500
RET-5.5_UGR
RET-5.5_DGR
Ret-5.5
rs7069590
1 kb
hg19
chr10:
43,582,000
43,582,500
43,583,000
RET+3_UGR
RET+3_DGR
RET+3
rs2435357
1100 bp
931bp
917bp
WT
RET -7(+/-)
1100 bp
953bp
961bp
WT
RET-5.5(+/-)
1100 bp
953bp
962bp
WT
RET+3(+/-)
Relative Expression (TaqMan qPCR)
1.2
1
0.8
0.6
0.4
0.2
0
**
**
**
Cas9 only Cells
RET+3(+/-)
RET-5.5(+/-)
RET-7(+/-)
RET
EDNRB
SOX10
GATA2
RARB
PAX3