THE REGULATORY LANDSCAPE OF THE *RET* GENE IN HIRSCHSPRUNG DISEASE

by
Kameko M Karasaki

A thesis submitted to Johns Hopkins University in conformity with the requirements for the degree of Master of Science

Baltimore, MD
April 2016

Abstract

Hirschsprung disease (HSCR), or congenital aganglionosis, is characterized by a contiguous lack of enteric neurons in variable segments of the gut. Both coding and non-coding mutations in the receptor tyrosine kinase *RET* are the major genetic drivers of the disease, although all disease causing mutations in *RET* have not yet been identified. A prior Genome Wide Association Study (GWAS) using 220 HSCR trios identified 38 common single nucleotide polymorphisms (SNPs) that are significantly associated with the disease and are located in non-coding regions at the *RET* locus. Eight of these SNPs disrupt known *RET* transcription factor binding sites in the human neuroblastoma cell line, SK-N-SH, and are located within enhancer elements. Three of these eight SNPs show differential enhancer activity between the non-risk and disease associated (risk) alleles, or an allelic difference. I sought to determine how many of the remaining 30 SNPs displayed enhancer activity and allelic differences by using a dual *in vitro* luciferase reporter assay in SK-N-SH cells. My studies revealed that 28 SNPs had enhancer activity while seven of those displayed allelic differences. The SNPs that showed an allelic difference and affected enhancer activity potentially disrupt binding sites for PAX3, MZF1, ZNF263, and Myb/Mybl1. Therefore, the genetic risk of HSCR susceptibility at *RET* is conferred by allelic differences at at least 10 enhancer elements, demonstrating the high degree of intra-locus risk heterogeneity.

Mentor/Principal Investigator: Aravinda Chakravarti, Ph.D.

Advisor: Robert D. Horner, Ph.D.

Acknowledgements

My deepest thanks to my mentor, Dr. Aravinda Chakravarti, for his guidance and support during my time in his lab. He has taught me to think critically and question everything, a tremendously important skill for any scientist.

Thanks to Dr. Sumantra Chatterjee for training me and assisting in the design and implementation of my thesis project. He spent countless hours helping me become confident in a wet lab setting and guiding me through my first independent study.

I would also like to thank the other members of the lab. Dr. Ashish Kapoor and Dr. Dongwon Lee contributed greatly to the data analysis in this study and I am extremely grateful for their time and patience. Maria X. Sosa and Dallas Auer have assisted and encouraged me every step of the way. They have each impacted my growth as a researcher and I wouldn't have been able to complete this project without their support.

Thank you to the Johns Hopkins University Department of Biology for giving me the opportunity to peruse my interests. Thanks especially to Dr. Kathryn Tifft for patiently guiding me through the process of completing a thesis. The education I have received here has been more than I had ever hoped for and has built an amazing foundation that I can't wait to build on as I continue my education.

Finally, I'd like to thank my family. I wouldn't be where I am today without the continued love and support of my parents, and I couldn't be more appreciative of all they've done for me.

Table of Contents

1. Introduction

Hirschsprung disease (HSCR), or congenital aganglionosis, is characterized by a failure of enteric neural crest-derived intestinal ganglion cells to fully migrate, proliferate, or differentiate through the length of the fetal intestines during development.[1] The lack of precursor cells leads to an absence of neurons in the myenteric and submucosal plexus along variable but contiguous lengths of the gastrointestinal tract. Without viable neurons, the affected length of bowel is non-functional and must be surgically resected. [1]

HSCR is classified into three distinct forms: short-segment (S-HSCR; aganglionosis up to the upper sigmoid colon), long-segment (L-HSCR; aganglionosis up to the splenic fixture and beyond), and total colonic aganglionosis (TCA; aganglionosis affecting the entirety of the large intestine). S-HSCR accounts for about 80% of cases, while L-HSCR and TCA account for 15% and 5% respectively.[1] Additionally, in less than 1% of cases, total intestinal agaglianosis (TIA; aganglionosis of the entire large intestine and also extending into the small intestine) occurs and is often lethal.[2,3] Since HSCR is a congenital disease, risk alleles and disease susceptibility are potentially evolutionarily maintained if carriers have a selective advantage. It is thought that HSCR and the associated slower bowel motility may provide a substantial survival advantage in communities where diarrheal illness was a major cause of death. [3] Otherwise, affected individuals would have been unlikely to survive to reproductive age before surgical intervention was possible, and therefore the risk alleles would not have remained in the gene pool.

HSCR is frequently associated with other malformation syndromes and chromosomal abnormalities (e.g. Waardenburg syndrome, Smith-Lemili-Opitz syndrome type II, trisomy 21, and interstitial deletion of chromosome 13q. [2,4] Hirschsprung Disease is the major genetic cause of functional intestinal obstruction in neonates with an incidence of ~1 per 5,000 live births. However, the incidence varies significantly among ethnic groups (1.5, 2.1, and 2.8 per 10,000 live births in Caucasians, African-Americans, and Asians, respectively). [1] There is also a marked sex bias and males are four times more likely than females to develop HSCR. [1]

During fetal development, the intestinal tract is innervated anteriorly to posteriorly due to a wave of morphogenetic changes. [5] The migration and differentiation of intramural ganglion cells can be attributed to a highly regulated and complex series of gene expression and interactions. The complex regulation of gut innervation is evident by the number of genes implicated in Hirschsprung disease. Previous studies have identified 14 genes with mutations in HSCR: *RET, GDNF, NRTN, SOX10, EDNRB, EDN3, ECE1, ZFHX1B, PHOX2B, TCF4, KIAA1279, GRFA1, PSPN, L1CAM*. [3,6] However, of these 14 genes, *RET* (a receptor tyrosine kinase, rearranged during transfection) is the major gene containing deficiency alleles, or alleles resulting in reduced expression, in >80% of cases. [6] Studies have shown that expression of HSCR-associated *RET* receptors lead to apoptosis, which may contribute to the absence of enteric neurons. [7]

HSCR can be either familial or sporadic. Approximately 20% of cases are familial, while the remaining 80% are sporadic. [2] Both forms may arise from a Mendelian mode of inheritance (both dominant and recessive with reduced

penetrance) as well as more complex, non-Mendelian modes of inheritance.[3] Mutations in *RET* are highly associated with the development of HSCR, but only 50% of familial and 15-20% of sporadic Hirschsprung cases have previously been reported to have causal mutations.[2] Less than 30% of HSCR cases overall have coding mutations in any of the genes known to be associated with the disease.[6] A contributing factor to the difficulty in identifying all genetic causes of HSCR is that generally, complex diseases, such as HSCR, are not caused by a single rare coding variant. Rather, a multitude of common variants with small effect sizes leads to the disease.[3] Multiple mutations having a compounding effect might also explain the varying degrees of severity found in HSCR.

However, because so many genes are associated with the disease, and many disease associated single nucleotide polymorphisms (SNPs) are present in and around these loci, a comprehensive map of all such regulatory SNPs is essential. A complete map of regulatory elements that are associated with HSCR would greatly increase the ability to diagnostically predict disease risk as well as contribute to a better understanding of how HSCR manifests during embryotic gut development. Since the majority of causal SNPs have been found in the *RET* gene, there is an ongoing effort by our lab to identify every causal SNP in the *RET* locus. Causal SNPs in non-coding DNA tend to lie within a regulatory region (such as an enhancer) and potentially disrupt important transcription factor (TF) binding sites, therefore dysregulating gene expression and hence function.

To identify SNPs associated with the development of HSCR, a Genome Wide Association Study (GWAS) was performed using 220 HSCR trios (an affected individual and both of their unaffected parents).[8] Thirty-eight common

SNPs (minor allele frequency >10%) were found at the *RET* locus and were associated with HSCR at a genome wide significance level ($p \leq 5x10^{-8}$). In addition to the associated SNPs, epigenomic data was also used to predict putative enhancer regions. DNaseI hypersensitivity, monomethylation of Histone 3 Lysine 4 (H3K4me1), and acetylation of Histone 3 Lysine 27 (H3K27ac) are all epigenetic markers that are a strong predictor of active enhancers.[9-11] The NIH Roadmap Project has compiled data for these marks in human fetal gut tissue, which were used in this study to mark putative gut enhancers.[12] Of the 38 SNPs, eight disrupt a known transcription factor (TF) binding site in a human neuroblastoma cell line, SK-N-SH.[13] The remaining 30 SNPs do not disrupt a known TF binding site but do overlap H3K4me1, H3K27ac, and/or DNaseI hypersensitivity regions in the fetal gut and, thus, may disrupt a yet uncharacterized biding site in an enhancer. These remaining 30 SNPs were assayed in an *in vitro* enhancer assay to determine first, if they are located in enhancer regions, and second, if the risk variant disrupts the function of the region. We found 28 of these SNPs are located in an enhancer element, and seven showed allelic differences in activity. Since all of the tested regions are present in and around the *RET* gene, and the SNPs are significantly associated with HSCR, it is highly likely they are regulatory elements controlling the expression of *RET*. We then looked at the 13 haplotypes (>1% frequency in the population) containing various risk and non-risk alleles for these 10 SNPS with allelic difference (seven from this study and three previously characterized) to ascertain variation in HSCR risk across variant haplotypes. Three of the haplotypes had an significant effect on risk of the disease compared to an almost

risk allele free haplotype and, furthermore, the haplotype containing all 10 risk alleles had an 11-fold increase in risk compared to control. In this study, we have generated a comprehensive functional map of all known disease associated SNPs at a locus for a complex disease and have shown that combinatorial effects from multiple enhancer variants are critical for a complete accounting of disease risk.

2. Materials and Methods

Polymerase Chain Reaction (PCR)

The UCSC Genome browser was used to obtain DNA sequence information for the region surrounding each SNP. DNA sequences approximately 500 bp long containing the SNP were identified (Appendix). Takara Clontech InFusion Primer Design and Primer3 were used to design primers with homology regions to the pGL4.23 [*luc2*/minP] plasmid (Appendix). All PCR reactions were 30 cycles long and used Phusion High-Fidelity DNA Polymerase (New England BioLabs Inc., Ipswitch, MA). All DNA samples came from the National Human Genome Research Institute's HapMap Project [14] (Coriell Cell Repositories, Camden, NJ) (Appendix). We carefully selected genomic DNA samples for each construct to minimize any additional variation between the non-risk and the risk allele since we wished to study each genetic variant in isolation. Following successful amplification, the PCR products were cleaned up by Qiaquick PCR purification Kit (Qiagen, Hilden, Germany).

Cloning

The PCR amplified region of interest was inserted by homology cloning into the cut pGL4.23 [*luc2*/minP] vector using an InFusion Cloning kit (Takara

Bio, Mountain View, CA). The InFusion Cloning protocol was followed. For this protocol, the desired vector was first linearized using a double digest. In this case, pGL4.23[*luc2*/minP] was linearized with KpnI-HF and XhoI (New England Biolabs, Ipswitch, MA). The primers designed for PCR amplifies the DNA sequence of interest, and includes 15 bp on each end of the PCR product that is homologous to the linearized ends of the vector. The InFusion enzyme recognizes the ends of the PCR products and the ends of the linearized vector and induces homologous recombination, resulting in a circular vector with the correct insert.

Following the cloning reaction, the vectors were transformed into Stellar Competent Cells (an E. coli HST08 strain). Transformed bacterial cells were plated on Luria Bertani (LB) agar plates containing ampicillin at 100µg/ml concentration. Cells were allowed to grow overnight, following which individual colonies were selected for a bacterial colony PCR reaction. The forward (CCAGTGCAAGTGCAGGTGCC) and reverse (CGTAGCGCTTCATGGCTTTG) primers used for the colony PCR are complementary to the vector backbone and can detect if a fragment of the right size has been inserted into the vector. If the colony was positive for an insert of the correct size, plasmid DNA was isolated from an overnight bacterial culture and sent for sequencing at the Johns Hopkins Genetics Research Core Facility using the same primers as the colony PCR reaction. Once sequencing confirmed the correct inserts, non-risk and risk allele sequences were aligned to the reference human genome assembly hg19/GRCh37 using Sequencher (Gene Codes Corporation, Ann Arbor, MI) to check for any differences between the cloned sequence and the human reference sequence. We also checked for any additional SNPs between the two constructs as well as any

additional variants that were not identified during the HAPMAP study (see Results).

Cell Culture

All assays were performed in the human neuroblastoma cell line SK-N-SH. The cells were grown in a media containing 500 ml DMEM (ThermoFisher Scientific, Waltham, MA), 50 ml Fetal Bovine Serum (ThermoFisher Scientific, Waltham, MA), 5 ml Pen-Strep (ThermoFisher Scientific, Waltham, MA), and 5 ml L-Glutamine (ThermoFisher Scientific, Waltham, MA). SK-N-SH, an adherent cell line, was grown in 10 cm dishes, and dissociated using 0.05% Trypsin-EDTA. (ThermoFisher Scientific, Waltham, MA).

Luciferase Assay

The non-risk and risk allele containing vectors were then used for dual luciferase reporter assays. SK-N-SH cells were plated on a 24 well plate, allowed to grow to about 50% confluence (typically overnight), and then transfected with 30 ng of a luciferase vector either containing the putative enhancer element or just the empty vector, and FuGene (a nonliposomal transfection reagent) (Promega, Madison, WI) mixed in Opti-MEM (ThermoFisher Scientific, Waltham, MA). Each well was also transfected with 6 ng of renilla luciferase vector as a transfection control. Transfected cells were allowed to grow to 90-100% confluency (24 hours). Each assay was performed in triplicate.

After 24 hours the media was removed and cells were washed with neutral phosphate buffered saline (PBS) before being lysed for 30 minutes by shaking at room temperature in 150 µl of Passive Lysis Buffer from a Dual Luciferase Reporter (DLR) Assay kit (Promega, Madison, WI). The lysate was spun down at

4˚C for 2 minutes to pellet any cellular debris, and then 30 µl of lysate from each well was plated in triplicate on a 96 well plate. A Tecan InFinite Pro workstation was used to perform the dual luciferase assay. The Dual-Luciferase Reporter Assay System protocol was followed, except each well was injected with 50 µl of LARII and Stop&Glo substrate for firefly and renilla luciferase respectively (Promega, Madison, WI).

Gene Expression Assay

RNA isolated from SK-N-SH cells was converted to cDNA using SuperScript IV Reverse Transcriptase (ThermoFisher Scientific, Waltham, MA). The SuperScript IV reverse transcription protocol was followed. The cDNA was then used with TaqMan Gene Expression assays to perform qPCR with probes Hs00992437_m1 (PAX3) and Hs01019337_m1 (ZNF263) with Human GAPD (GAPDH) Endogenous Control as an internal loading control (ThermoFisher Scientific, Waltham, MA).

Haplotype Odds Risk Determination

Once the SNPs displaying an allelic difference were determined, haplotypes present in a control (non-HSCR) and case (HSCR) population were created. Control population haplotype data was collected from the CEU, FIN, GBR, IBS, and TSI (total of 503 European ancestry samples) populations in the 1000 Genomes Project (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/).[15] Case population haplotype data was collected from the 220 HSCR Trio probands from the HSCR GWAS[8] using Affy500K[8] and HapMap imputation[14]. The haplotypes were analyzed with Haploview[16] and Plink

(http://pngu.mgh.harvard.edu/purcell/plink).[17] Standard methods were used for calculation of odds ratios (OR), upper and lower confidence limits and significance of deviation from the null hypothesis of no association (OR = 1) by a χ^2 statistic.[18] The odds of one risk allele (CTGAACCACT) was set to 1 as there were no individuals with a haplotype containing no risk alleles.

3. Results

Non-Risk and Risk Vector Cloning

The SNP IDs used throughout this study was assigned by NCBI's dbSNP (http://www.ncbi.nlm.nih.gov/SNP). After sequencing, the risk and non-risk vector sequences were aligned using Sequencher to ensure the correct allele was cloned. The non-risk and risk alleles are shown in Table 1. Some constructs had additional SNPs that were different in the sequences around the SNP under study despite efforts to minimize the background SNPs. One pair of constructs had an additional SNP not reported in HapMap.

The locations of all 38 HSCR associated SNPs in this study are shown in Figure 1. The SNPs lay upstream and in the first intron of *RET*. The eight SNPs that disrupted known TF binding sites in SK-N-SH cells are shown in Figure 2. All eight are in regions that display enhancer activity, while three also showed an allelic difference.

Table 1: Non-Risk and Risk Alleles for each polymorphism tested

SNP	SNP ID	Non-Risk Allele	Risk Allele	Additional SNPs different between the constructs	Additional SNPs different between the constructs and reference
1	rs788267	C	T		
2	rs788263	C	G	rs1539291	
3	rs788262	A	G		
4	rs788261	T	C		
5	rs788260	G	A	rs186489714	
6	rs2995411	C	T		
7	rs1582227	T	C		
8	rs2488278	T	C	rs77947964	
10.11	rs17158318/ rs17158320	G, C	A, A		
12	rs947696	G	T	rs1815722	
13	rs7908085	T	A	rs2185791	
14	rs10900290	C	T		
15	rs947690	G	C		
16	rs1547930	G	A		
17	rs3004258	T	G		
20	rs3026703	T	C		
21	rs3026707	A	G	rs3026706, rs3026709, rs4948560	
22	rs2505989	G	C		
24	rs741763	G	C		
25	rs2505995	A	G		
26	rs2506010	C	T		rs3123717
28	rs2506020	C	T	rs2506019, rs2435361, rs28409950, rs28576188, rs28464869	
30.31	rs2506021/ rs2435342	C, T	T, C		
32	rs2506022	C	T		
33	rs2435343	T	G		
34	rs12247456	G	A		
36	rs7393733	G	C		
38	rs2505541	T	C		

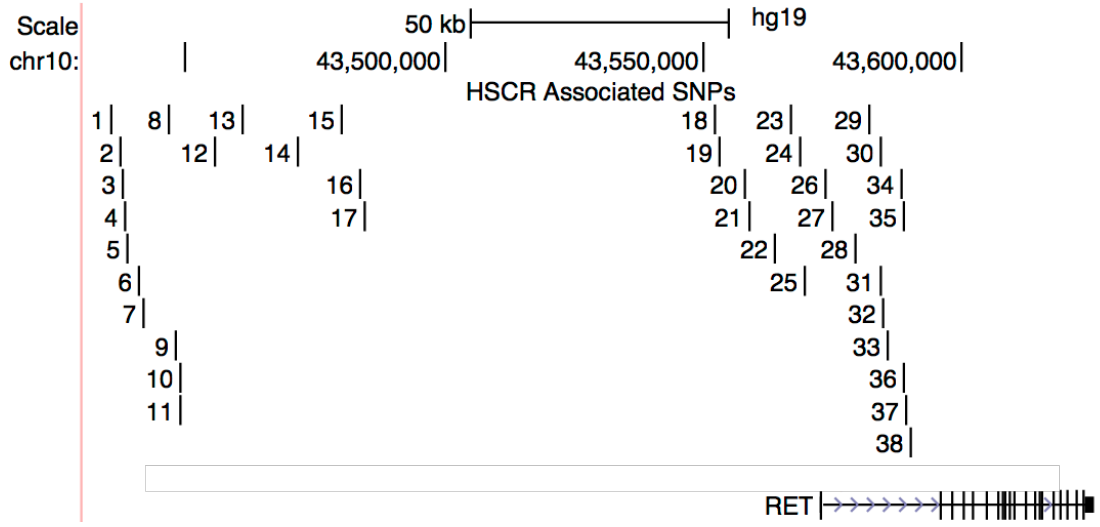

Figure 1: 200 kb (chr10:43,430,437-43,629,595, human genome assembly 19) region around the *RET* gene showing all 38 HSCR-associated SNPs. SNP IDs are provided in Table 1.

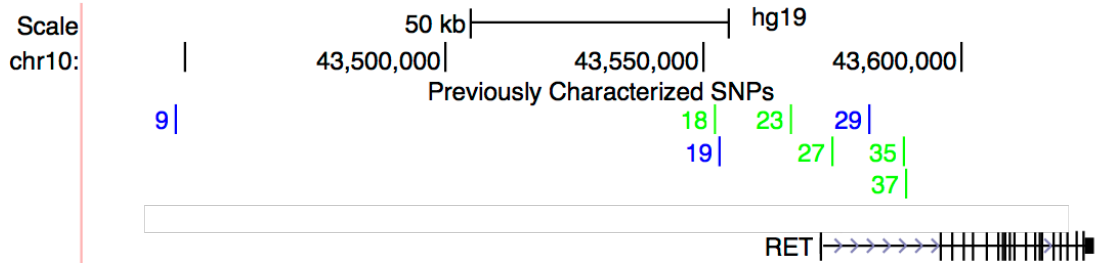

Figure 2: 200 kb (chr10:43,430,437-43,629,595) region around the *RET* gene showing previously tested SNPs. SNPs colored green or blue are part of enhancer elements. The SNPs colored blue showed significant allelic difference between the non-risk and risk alleles.

Enhancer Function Determination

Luciferase assays with ~500bp regions containing the 30 disease associated SNPs showed that 28 had enhancer activity, defined as >2 fold activity as compared to the empty vector (Figure 3). Among these 28 SNPs displaying enhancer activity, seven SNPs also showed a significant gene expression difference driven by the risk allele as compared to the non-risk allele. In

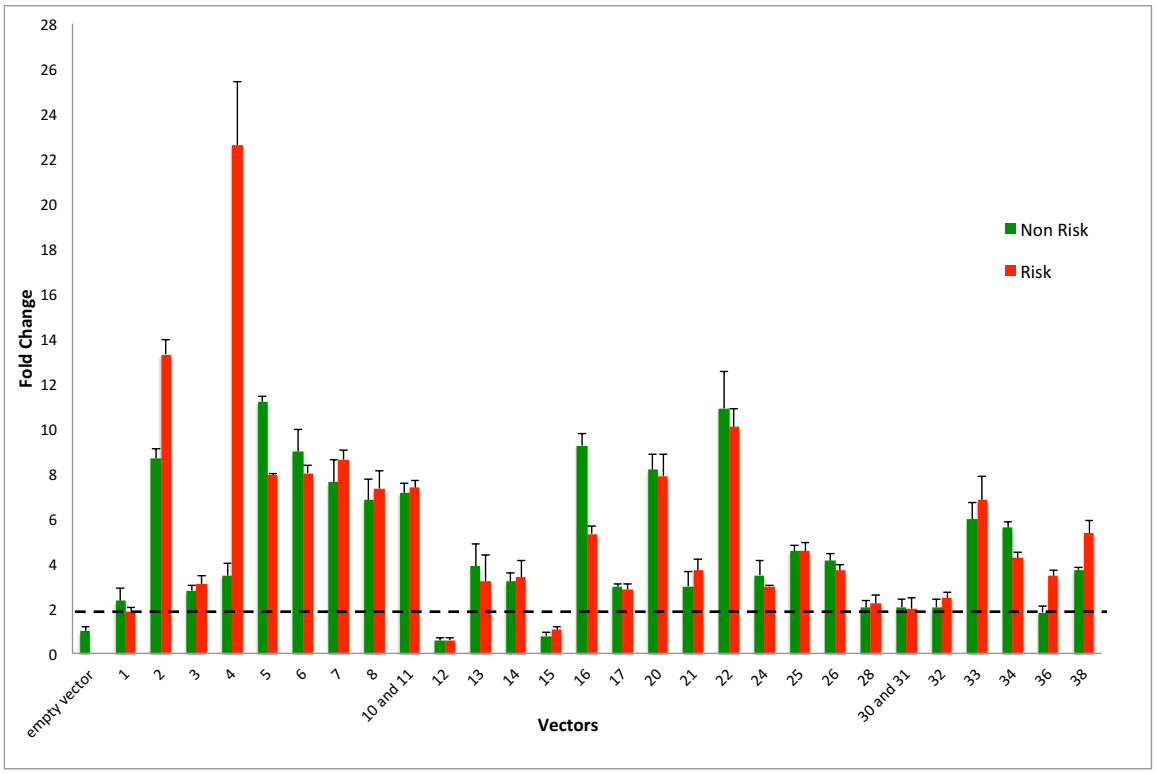

Figure 3: Luciferase assay activity for 38 HSCR-associated SNPs. Data are reported as fold change relative to expression of luciferase driven by only minimal promoter vector (pGL4.23 empty). All regions that drive expression >2 fold as compared to empty vector are considered an enhancer (black serrated line). Error bars represent the standard error of the mean.

combination with the eight previously characterized SNPs, 34 SNPS are in regions that showed enhancer activity and 10 had a significant allelic difference (Figure 4 and Figure 5). The risk alleles of SNPs 5, 9, 16, 19, 29, and 34 showed reduced enhancer ability compared to the non-risk allele, while the risk allele of SNPs 2, 4, 36, and 38 showed increased enhancer activity compared to the non-risk allele. Thus, the results of this current study increased the number of regulatory elements in this locus as well as the SNPs which disrupt enhancer activity and might have a role in controlling the expression of *RET*. We now have a more detailed view of the regulatory universe of *RET* and the causal SNPs in the region which are associated with HSCR.

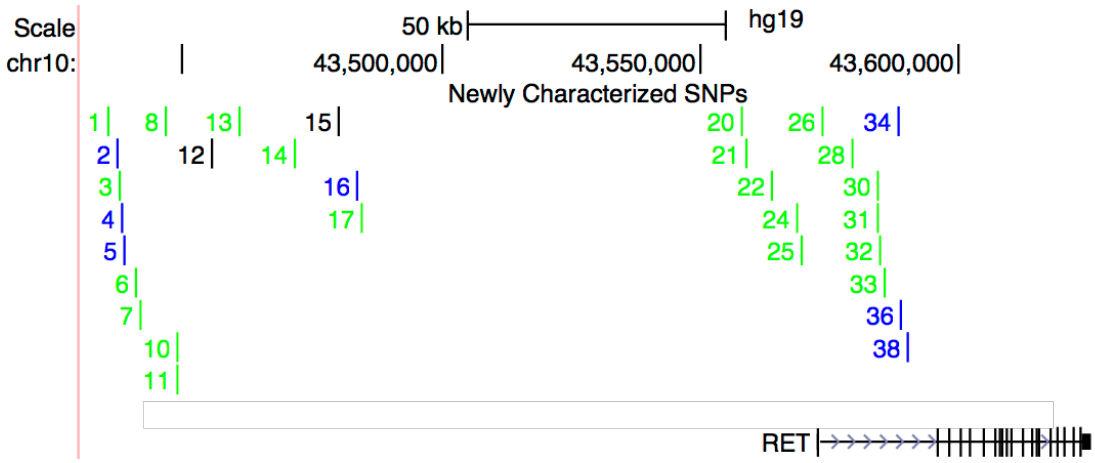

Figure 4: 200 kb (chr10:43,430,437-43,629,595) region around the *RET* gene highlighting 30 SNPs tested in this study . SNPs shown in black are not part of an enhancer element. SNPs colored green showed enhancer activity in *in vitro* luciferase assays and SNPs colored blue showed significant allelic difference between the non-risk and risk alleles.

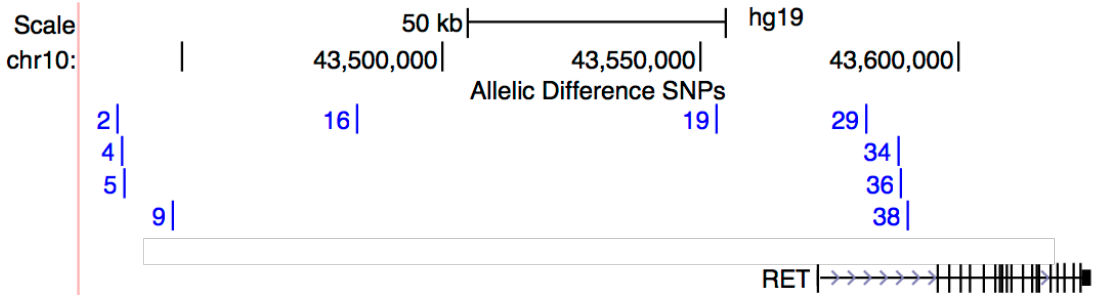

Figure 5: All SNPs (both previously and newly characterized) displaying an allelic difference in a 200 kb (chr10:43,430,437-43,629,595) region around the *RET* gene.

Potential Transcription Factor Binding Disruption

Regulatory elements, such as transcriptional enhancers, bind various transcription factors (TFs) for their activity. To determine the potential TFs that bind to enhancers that display an allelic difference we used JASPAR [19] and TRANSFAC[20], transcription factor binding site databases. A 20 bp region surrounding the SNP was compared to ~900 known TF position weight matrices (PWMs). TF PWMs are weighted measures of experimentally determined TF binding sequences, which allows us to potentially predict which TF can bind a given DNA sequence in the absence of experimental data on that region for that cell line.

The most significant matches we detected were for PAX3, MZF1, ZNF263, and Myb/Mybl1 (Table 2). PAX3, MZF1, and ZNF263 bind to sequences containing the non-risk allele of SNP 36. ZNF263 is also predicted to bind a sequence containing the non-risk allele of SNP 16. However, the PAX3 and MZF1 predictions are based on few experimentally determined TF binding sequences

and therefore need to be verified by another method to ascertain significance; moreover, zinc finger binding domains are typically GC-rich sequences and therefore give high statistical significance owing to their rarity. Myb/Mybl1 are predicted to match the risk allele of SNP 2, which may contribute to the increased enhancer activity observed for the risk allele. There is also an ETS1 binding motif around the risk allele SNP 36 and a MA0055.1 binding motif around the non-risk allele of SNP 38, but the SNP does not disrupt the core sequence for either. Although it is possible that a mutation outside of the core sequence could disrupt binding, it is not as likely to change TF binding as a mutation within the core sequence is. MA0080.2 (another ETS TF) shows equal binding potential to both the non-risk and risk allele of SNP 36. ZID shows equal binding potential to both alleles of SNP 5. The Srf_secondary motif at the non-risk allele of SNP 4 doesn't seem to be a real binding site since it is a secondary motif and the UA8 motif that matches both the risk and non-risk allele of SNP 4 belongs to an uncharacterized TF so further analysis (e.g. TF ChIP-seq) would need to be performed to draw any conclusions about that binding motif. We also detected a non-canonical binding motif for ATF1 near SNP 36. Bioinformatic analysis did not predict any potential binding site around SNP 34.

One way to determine if a TF could bind to an enhancer element is to determine if the TF is present in the cell line used for luciferase assays (SK-N-SH). If the TF is not present in the cell, then it is unlikely that that the TF is affecting the enhancer activity. A gene expression assay was performed on two TFs: *PAX3* was selected based on its known association with *RET*[20], and *ZNF263* was selected based on its match with two of the SNPs displaying an allelic

Table 2: Position weight matrix (PWM) of various transcription factor classes

SNP #	Allele	TF	Motif	Site at Locus	Start	Stop	Strand	Log-odds	p-value
2	Risk	Myb	Myb_primary	CAATAAACCGTTAAGTA	2	18	+	12.3	$3.72*10^{-5}$
		Mybl1	Mybl1_primary	CAATAAACCGTTAAGTA	2	18	+	13.1	$2.12*10^{-5}$
	Non-Risk	PAX2	V_PAX2_02	AATAAACCC	3	11	+	8.9	$4.67*10^{-5}$
4	Risk	unknown	UA8	TGAGAAGAAACAGC	4	17	-	11.3	$8.68^{*}10^{-5}$
		AREB6	V_AREB6_04	CTGTTTCTT	5	13	+	11.0	$8.37*10^{-5}$
	Non-Risk	SRF	Srf_secondary	CTGAGAAAAAACAGCCT	2	18	-	10.2	$5.16*10^{-5}$
		SRF	Srf_secondary	CCTGAGAAAAAACAGCC	3	19	-	10.6	$3.45*10^{-5}$
		unknown	UA8	TGAGAAAAAACAGC	4	17	-	12.8	$2.95*10^{-5}$
5	Risk	ZBTB6	V_ZID_01	TGGCTCTATCATG	5	17	-	15.7	$2.74*10^{-6}$
	Non-Risk	ZBTB6	V_ZID_01	TGGCTCCATCATG	5	17	-	15.9	$2.38*10^{-6}$
16	Non-Risk	ZNF263	ZNF263	GCCTCACTGCTCCAG	3	17	-	9.7	$8.09*10^{-5}$
36	Risk	SPI1	MA0080.2	AGGAAGT	13	19	+	12.7	$6.51*10^{-5}$
		Atf1	Atf1_secondary	GAGTGACGAGGAAG	5	18	+	13.3	$4.68*10^{-6}$
	Non-Risk	PAX3	V_PAX3_01	TCGTCACTCTTAC	1	13	-	5.5	$9.99*10^{-5}$
		MZF1	V_MZF1_02	GAGTGAGGAGGAA	5	17	+	12.1	$3.28*10^{-5}$
		SPI1	MA0080.2	AGGAAGT	13	19	+	12.7	$6.51*10^{-5}$
		ETS1	ETS1	CACTTCCTCCT	10	20	-	13.5	$1.05*10^{-5}$
		ZNF263	ZNF263	CACTTCCTCCTCACT	6	20	-	10.2	$6.72*10^{-5}$
38	Non-Risk	E-box	MA0055.1	GGAGAGCTGGTG	1	12	+	9.7	$3.46*10^{-5}$

The log-odds shown above used a cutoff of (stringent) p value $<10^{-4}$

difference. Gene expression is measured using cycle threshold (Ct) values determined by the number of qPCR cycles required for the fluorescent signal of the gene of interest to be detected. Ct values for the gene of interest are compared to Ct values for a housekeeping gene. The comparison is called the ΔCt; the lower the ΔCt, the higher the gene expression and vice versa. Based on the ΔCt values (normalized expression compared to GAPDH gene), *PAX3* is moderately expressed and therefore could potentially have its binding disrupted by the risk allele of SNP 36. However, the expression of *ZNF263* is very low or absent in this

cell line. As noted earlier, all zinc finger TFs have a strong bias for GC rich motifs so it is possible that another zinc finger family protein binds instead of ZNF263 (Table 3 and Figure 6).

Table 3: Ct Values for GAPDH, PAX3, and ZNF263 in SK-N-SH cells

	GAPDH (Ct)	PAX3 (Ct)	ZNF263 (Ct)
Assay 1	27.99	29.13	33.99
	27.56	28.89	34.28
	27.90	29.40	34.61
Assay 2	27.55	29.15	34.31
	27.68	28.97	34.22
	27.83	29.18	34.32
Assay 3	28.98	28.94	33.27
	28.14	29.00	33.43
	28.40	28.70	33.38
Assay 4	27.54	28.92	33.59
	28.88	28.71	33.05
	28.89	28.67	33.30

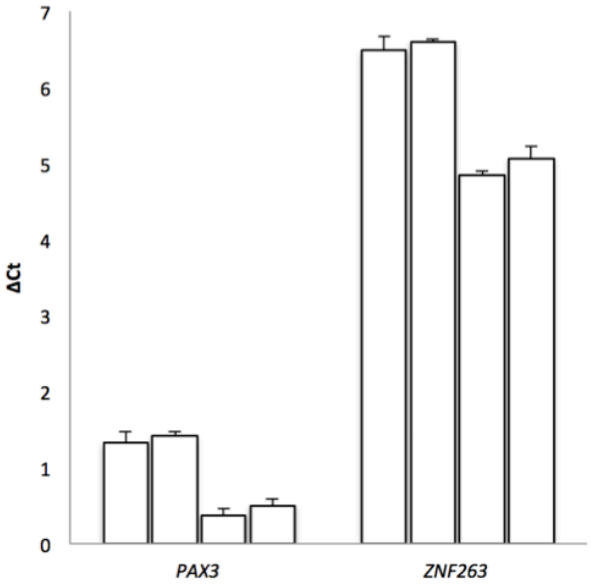

Figure 6: Ct data used for ΔCt calculation and ΔCt for two implicated TFs (from Jaspar and Transfac analysis). TFs PAX3 and ZNF263 transcript levels are normalized to the GAPDH levels. Error bars represent the standard error of the mean.

Evolutionary Conservation

Evolutionary conservation was an important feature for identifying enhancers in the past; previous studies aimed at identifying enhancers in the RET locus mapped putative enhancers solely via conservation [16]. However, more recent studies have shifted away from using evolutionary conservation and instead focus on epigenetic modifications to inform their enhancer predictions. For example, the 38 SNPs tested in this study were identified partially due to

their epigenetic modification marks. Therefore, in order to determine how the new data collected during this study compared to previous knowledge of regulation at the *RET* locus, we also mapped all evolutionary conserved regions (70% sequence identity, ≥100 bp length) from 12 non-human vertebrates (chimpanzee, baboon, cow, pig, cat, dog, rat, mouse, chicken, zebrafish, *Fugu*, and *Tetradon*) using an updated conservation assembly to see which of the 38 HSCR associated SNPs determined by a GWAS [8] overlapped a conserved region (Figure 7). Of the 38 SNPs that were tested for this project, only 14 were found in regions previously determined to be enhancers based only on sequence conservation [21] (Table 4). Of these 14 SNPs, 13 proved to have enhancer activity and six of them also displayed an allelic difference between the ancestral (non-risk) and disease associated (risk) alleles. Eight additional SNPs were shown to be part of a conserved region in the new conservation analysis but were not tested previously (rs788263, rs788260, rs2488278, rs947696, rs10900290, rs3026703, rs3026707, rs2505995). Of these SNPs, seven showed enhancer activity and two showed allelic differences. In total, evolutionary conservation alone is able to detect 20 out of the 36 regions that displayed enhancer activity in this study, and eight of the 10 SNPs that show an allelic difference. However, of the remaining 16 SNPs not mapping to previously reported enhancer regions or ECR, 15 had enhancer activity and two displayed an allelic difference. Thus, sequence conservation coupled with the use of enhancer marks (H3K4me1, H3K27ac, and DNase Hypersentivity) and disease associated variants can be powerful tools in discovering regulatory regions in the genome and creating a more complete genomic map.

Table 4: SNP location compared to Previously Tested Enhancers (PTE) and Evolutionarily Conserved Regions (ECR)

SNP #	SNP ID	Overlap PTE	Overlap ECR	Enhancer Ability in this Study	Allelic Difference
1	rs788267	No	No	Yes	No
2	rs788263	No	Yes	Yes	Yes
3	rs788262	No	No	Yes	No
4	rs788261	No	No	Yes	Yes
5	rs788260	No	Yes	Yes	Yes
6	rs2995411	No	No	Yes	No
7	rs1582227	No	No	Yes	No
8	rs2488278	No	Yes	Yes	No
9	rs2506030	Yes	Yes	Yes	Yes
10	rs17158318	No	No	Yes	No
11	rs17158320	No	No	Yes	No
12	rs947696	No	Yes	No	No
13	rs7908085	No	No	Yes	No
14	rs10900290	No	Yes	Yes	No
15	rs947690	Yes	Yes	No	No
16	rs1547930	Yes	Yes	Yes	Yes
17	rs3004258	No	No	Yes	No
18	rs4948702	Yes	No	Yes	No
19	rs7069590	Yes	Yes	Yes	Yes
20	rs3026703	No	Yes	Yes	No
21	rs3026707	No	Yes	Yes	No
22	rs2505989	Yes	Yes	Yes	No
23	rs2435367	Yes	No	Yes	No
24	rs741763	No	No	Yes	No
25	rs2505995	No	Yes	Yes	No
26	rs2506010	No	No	Yes	No
27	rs2506011	Yes	Yes	Yes	No
28	rs2506020	No	No	Yes	No
29	rs2435357	Yes	Yes	Yes	Yes
30	rs2506021	No	No	Yes	No
31	rs2435342	No	No	Yes	No
32	rs2506022	No	No	Yes	No
33	rs2435343	No	No	Yes	No
34	rs12247456	Yes	Yes	Yes	Yes
35	rs752978	Yes	Yes	Yes	No
36	rs7393733	Yes	Yes	Yes	Yes
37	rs2506024	Yes	No	Yes	No
38	rs2505541	Yes	No	Yes	Yes

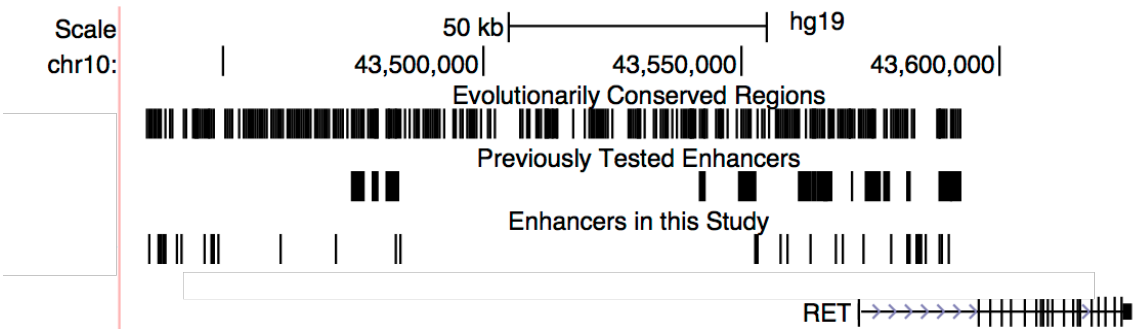

Figure 7: Evolutionary Conservation at the *RET* locus (chr10:43,430,437-43,629,595). The ECR (evolutionarily conserved regions) track shows regions with at least 70% sequence identity over 100 bp across 12 non-human vertebrates. DNA elements that acted as enhancers in previous studies were tested based on ECR data are shown in the "Previously Tested Enhancers" track. DNA elements that acted as enhancers in this study are based on the ~500 bp regions surrounding the HSCR associated SNPs and are shown in the "Enhancers in this Study" track.

Haplotype frequencies in control and case populations

In order to determine how a specific haplotype affects an individual's chance of disease, the odds ratio was calculated. The odds ratio is the ratio of the frequency of a haplotype in a population with the disease compared to frequency of the haplotype in a control population of the same ethnicity. (Table 5). An odds ratio greater than 1 indicates that individuals with the haplotype are more likely to have the disease, while an odds ratio of less than 1 indicates that the haplotype is protective. The haplotype with one risk allele in it is set to unity (1) and odds ratio for other haplotypes were calculated relative to it. Usually the haplotype with no risk alleles in it is set to unity, but in this case we selected only haplotypes which were present with >1% frequency; the haplotype with no risk allele was not

observed. Because of the sample size (220 individuals in the disease population, 503 individuals in the control population) and varying associations between the SNPs, some of the haplotypes aren't present at all or are present at a very low frequency, and, therefore, it is impossible to draw any meaningful conclusions about these haplotypes. Our data showed that three haplotypes have a statistically significant odds ratio: CTGAGCCGGC (OR: 0.36, P=0.002), CTGAGTTGGT (OR: 5.89, P= 2.76×10^{-5}), and GCAGGTTGGT (OR: 11.01, P=1.21×10^{-9}) (SNP order 2, 4, 5, 9, 16, 19, 29, 34, 36, 38 with green indicating non-risk allele and red indicating risk allele). Understanding the odds ratio

Table 5: Case (HSCR) and control population haplotype data.

Haplotype	Freq in Cases	Freq in Controls	Odds Ratio	P-value
CTGAACCACT	0.02	0.06	1.00	1
CTGAATCACT	0.03	0.05	2.01	0.246
CTGAGCCACT	0.03	0.06	1.50	0.078
CTGAGTCACT	0.03	0.05	2.41	0.427
CTGAATCGGC	0.06	0.11	1.94	0.147
CTGAGCCGGC	0.01	0.05	**0.36**	0.002
CTGAGTCGGC	0.05	0.14	1.34	0.512
CTGAATTGGT	0.02	0.02	3.76	0.022
CTGAGTTGGT	0.12	0.07	**5.89**	2.76×10^{-5}
GCAGGCCACT	0.01	0.05	0.81	0.730
GCAGGTCACT	0.03	0.06	1.84	0.204
GCAGGTCGGC	0.06	0.10	1.86	0.160
GCAGGTTGGT	0.54	0.17	**11.01**	1.21×10^{-9}

In the haplotypes, the alleles appear in the following order: 2, 4, 5, 9, 16, 19, 29, 34, 36, 38. Non-risk alleles are shown in green and risk alleles are shown in red. Odds ratios with significant P-values are highlighted in yellow.

associated with a specific haplotype can greatly improve the ability to understand the combinatorial effect of these non-coding variants. Additionally, determining if the alleles are statistically associated or not (and to what degree) can help us to impute the genotype of one SNP based on the genotype of another. We also note

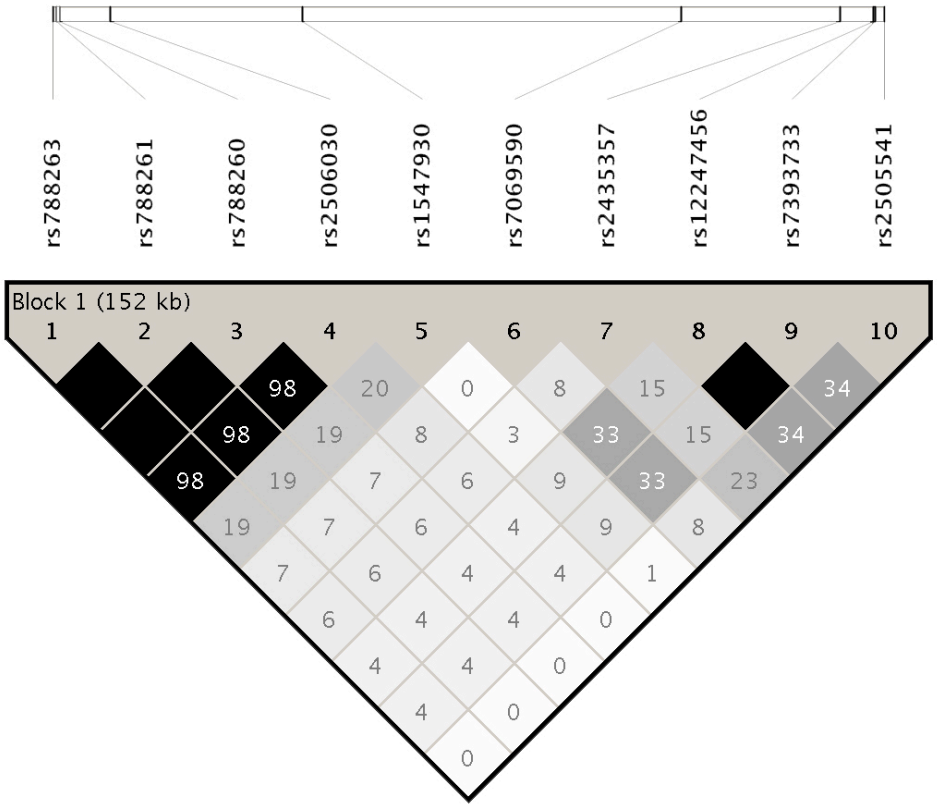

Figure 8: Linkage Disequilibrium Map of the ten SNPs displaying an allelic difference, shown in the following order: 2, 4, 5, 9, 16, 19, 29, 34, 36, 38. The bar at the top of the figure shows the positions of the SNPs relative to each other with the SNP IDs listed underneath. The R-squared values are shown at the diagonal intersection of the two SNPs. Black squares without an R-squared value indicate a value of 100, or that the two SNPs are in perfect linkage disequilibrium.

that two alleles can be in high LD can but can be part of two independent enhancers and also show allelic difference in expression. For example, rs788263, rs788261, and rs788260 are in complete linkage disequilibrium based on the samples we analyzed, but are likely independent factors (Figure 8). However, when the SNPs were tested in individual vector constructs, they each showed a distinct effect on gene expression, indicating that although they are completely associated there is a combinatorial effect to their linkage and not a single SNP responsible for the entirety of the change in gene expression.

4. Discussion

There are many susceptibility genes that add to the multifactorial risk of HSCR.[3] Nevertheless, despite considerable genetic heterogeneity in HSCR, *RET* is the most critical disease gene since deleterious coding variants occur in 21%, intragenic deletions in 5%, and enhancer variants in >98% of HSCR patients.[22] When examining the haplotypes present in control and case populations in this study, it becomes evident that most humans carry a non-coding *RET* deficiency allele since no individuals with all non-risk alleles were identified (Table 5). Furthermore, 17% of the individuals in the control population (i.e. individuals that do not have HSCR) have a haplotype that consists entirely of HSCR-associated risk alleles. Despite harboring a haplotype increasing the odds of developing HSCR 11-fold, it is still possible for individuals to be unaffected. Therefore, *RET* deficiency per se is necessary, not sufficient, for disease onset.

In the recent past, thousands of genetic variants affecting hundreds of traits and diseases have been discovered through GWAS.[23] Despite widespread

non-coding polymorphisms, multiple enhancers controlling each gene, and enrichment of disease association signals within such enhancers, there has been a distint lack of effort to do a comprehensive functional screen of all such variants even at a single locus. Therefore, we still do not know how many enhacers a gene can have, how many affect expression in a specific cell type, how many of these enhancers are mutant in human disease and how do they lead to disease given their individual weak effects.

Of a comprehensive functional assay of 30 HSCR associated SNPs, we found 28 of them are part of distinct loci with enhancer ability. Furthermore, seven of those SNPs exhibited an allelic difference between the non-risk and risk alleles and therefore potentially disrupt *RET* transcription and contribute to the development of HSCR. Testing all of the variants individually allowed us to construct haplotypes containing alleles displaying disruption of enhancer function. From this, we were able to calculate disease risk for various haplotypes, thus highlighting the importance of the combinatorial effect of the individual variants on the disease risk since none of the identified SNPs cause the disease on their own.

This study also determined several transcription factors are candidate regulators of some of the enhancers at *RET* and whose binding might be disrupted by the SNPs displaying an allelic difference. The identification of putative TF binding sites opens the possibility of understanding the regulation of *RET in vivo* in a more detailed manner, including its upstream regulators as well as allowing us to build gene regulatory networks (GRNs) containing RET and its associated functional partners, which can explain how the disruption of an

enhancer by a SNP might have a larger than expected effect on gene expression due to disruption of the GRN as a whole.

It is important to note that some of the SNPs tested in this study are clustered in such a way that they could be part of the same enhancer element or a part of a super enhancer.[17] Additionally, the combinatorial effects of the enhancers may not be fully appreciated. *In vivo,* enhancer elements act dynamically and combinatorially and are generally within large regulatory domains called topologically associated domains (TADs). [25] By testing the regions individually, any *in vivo* associations within a TAD would have been eliminated. Thus, a complete understanding of the genetic effect of a non-coding variant on a gene or disease requires knowledge of its multiple enhancers, how the target gene interacts with these enhancers, and if the SNPs cause a disruption of its GRN.

Identifying all causal SNPs is imperative to understand the complexity of a multifactorial disorder, and it may also give us mechanistic insights in to the progression of the disease. Combining knowledge from forward genetic screens with functional validation of the role of various TFs will help in better predicting the consequences of various non-coding variants associated with a disease

Limitations

There are a few limitations that are important to consider regarding the results of this study. The first is that while the *in vitro* luciferase assay does test for enhancer activity, there is no direct proof that these are enhancers of *RET*. In humans, enhancers can be either upstream or downstream of the target gene, and they can also be 1 Mb or more distal. [26] So, even if a construct DNA segment is shown to be an enhancer in a heterologous assay, there is no guarantee that it

regulates *RET* or that SNPs with differential allelic activity leads to effect on the transcription of the gene. These regions of DNA will need to be evaluated by either deleting or creating the exact variant in a model organism to ascertain how they impact the transcription of *RET,* and in turn the differentiation and proliferation of enteric neurons.

The second limitation is about the nature of the control experiment we utilized. When fragments of DNA are inserted into the pGL4.23 [*luc*2/minP] vector, the assay can determine if there are any sequences within the larger fragment that can drive gene expression. However, this is being compared to an empty vector, or a pGL4.23 vector with nothing inserted into its multiple cloning site. Because of this, the empty vector is ~500 bp smaller than vectors with an insert. This represents a ~10% size change with an enhancer fragment inserted. It is quite likely that any piece of DNA can drive some transcription so that an effective control is difficult to define for the cloning step. However, it is impossible to determine if the size of the vector with the insert changes anything compared to the empty vector such as transfection efficacy, etc.

Finally, there are always limitations when using cell culture to model *in vivo* cellular functions. SK-N-SH human neuroblastoma cells were used for these assays because a human neuronal cell line likely has somewhat similar gene expression patterns compared with human fetal enteric neuron cells. However, fetal cells and adult cells are markedly different developmentally and it is possible for fetal gut cells to express TFs that SK-N-SH cells do not, or vice versa. The transcription machinery in the SK-N-SH cells are responsible for all of the

data for this study, so considerations should be made when generalizing the results to *in vivo* conditions.

Future Directions

Since enhancers *in vivo* work dynamically and combinatorially, the SNPs need to be tested together in order to more closely model their actual effect on gene expression *in vitro*. A multiple mutant construct could be made to mirror the haplotypes found in the human population and then used in gene expression assays similar to those used in this study. Although this would still suffer from the limitations associated with *in vitro* experiments, it would create a more complete picture of the total effect of the enhancer regions.

Causative SNPs most likely disrupt TF binding sites and therefore affect gene expression. Although TF binding motif databases exist, they are predictive and require further experimental validation. The best way to do this would be to perform a TF-ChIP experiment for various putative TFs determined to bind sequence prediciton at the *RET* locus *in vivo* (in this case, in human fetal gut cells). This would eliminate the uncertainty of whether or not a TF is present in the cells at the time of development as well as the uncertainty of the TF binding domain.

Conclusion

HSCR is a complex disease with multifactorial inheritance. Despite its relatively high occurrence rate, the genetic basis of HSCR is still not completely identified. Mutations in *RET* have been implicated as the major genetic cause of HSCR. This study examined 38 SNPs associated with HSCR and identified 36 putative enhancer regions and 10 potentially causative SNPs at the *RET* locus.

Identifying all causative SNPs will greatly enhance the predictive power of genetic tests for HSCR as well as advance an understanding of how the disease develops. The putative enhancer elements and potentially causative SNPs identified in this study contribute to the development of a complete regulatory map of the *RET* locus and may help with the diagnosis and treatment of HSCR.

5. Appendix

Appendix Table 1: ~500 bp regions of DNA surrounding each SNP are shown. Rows that are greyed out are SNPs that were previously characterized and therefore not included in this study. The "Start" and "Stop" columns refer to the chromosomal coordinates where the fragment surrounding the SNP begins and ends. All 38 SNPs are found in the *RET* locus on chromosome 10.

SNP #	SNP ID	Start	Stop
1	rs788267	43434932	43435936
2	rs788263	43437007	43437506
3	rs788262	43437440	43437943
4	rs788261	43437726	43438225
5	rs788260	43438228	43438727
6	rs2995411	43440554	43441053
7	rs1582227	43441446	43441952
8	rs2488278	43446082	43446581
9	rs2506030		
10	rs17158318	43448581	43449090
11	rs17158320	* included with SNP 10 in a single vector	
12	rs947696	43455083	43455582
13	rs7908085	43460567	43461065
14	rs10900290	43471322	43471823
15	rs947690	43479479	43479979
16	rs1547930	43483056	43483559
17	rs3004258	43483900	43484403
18	rs4948702		
19	rs7069590		
20	rs3026703	43557546	43558048
21	rs3026707	43558368	43559122
22	rs2505989	43563195	43563700
23	rs2435367		
24	rs741763	43568087	43568586
25	rs2505995	43569379	43569878
26	rs2506010	43573167	43574025
27	rs2506011		
28	rs2506020	43578754	43579373
29	rs2435357		
30	rs2506021	43583869	43584408
31	rs2435342	* included with SNP 30 in a single vector	
32	rs2506022	43584264	43584845
33	rs2435343	43585384	43585889
34	rs12247456	43587982	43588420
35	rs752978		
36	rs7393733	43588440	43588915
37	rs2506024		
38	rs2505541	43589862	43590368

Appendix Table 2: Forward and Reverse primer sequences for the ~500 bp regions surrounding the SNPs listed in Appendix Table 1. The primers shaded in grey are primers that I already had from a previous project, and were cloned by double cutting with KpnI-HF and SacI-HF in CutSmart Buffer and ligating with T4 DNA Ligase rather than homologous recombination. TGGCCTAACTGGCCG was added to the 5' end of all forward primers as the 15 bp homology arm and TCTTGATATCCTCGA was added to the 5' end of all reverse primers as the 15 bp homology arm.

SNP #	SNP ID	Forward Primer (5'-3')	Reverse Primer (5'-3')
1	rs788267	CCAATAGGCAGGCAAGACAC	GACCCACTGAAGTCTGGAGG
2	rs788263	CACTCTCGCTCCCTCATTTG	CCCTCAGAGTCTGGGCAGT
3	rs788262	GAACCTCCCTAGCCAGTTCC	GTTCCTCATCCTGCTCCTGT
4	rs788261	GCACATGCATGGAGAACTGT	CTCTGCTGGGCTATCTGGAG
5	rs788260	GCGGAGACAAGATGATCCA	TGCTGAAAAGCCTCTTTGGT
6	rs2995411	AGGCAGGAGAATAGCGTGAA	TGCTATGACTTCTTGGTGGGT
7	rs1582227	CTCCTGACTTCGTGATCTGC	TCCCTCTTCTGTGATGTAGGC
8	rs2488278	CTGCCGGTTCTTGGTTTTT	GGCTCCTCCTTGGTGCTTAT
10	rs17158318	TGCTCTACCATTTCCGCAGA	ACACACAGCATCCTCTCTCC
11	rs17158320	* included with SNP 10 in a single vector	
12	rs947696	AGGTACCTAGCTCACCATGT	ATAGCTGGGATCATAGGCCC
13	rs7908085	TCTTTCATTCCAATCTGCATGTC	TCAGCAAATCAAATTCAGCAACA
14	rs10900290	CCTGATGACCCCTGCTTTCA	TGCCTCAATCTTACCTGCCA
15	rs947690	TCCTGTGGTATGATGGTGCA	ACGGGAGCAATGTTCTGGAG
16	rs1547930	AGCTCCCTGAAGTCACGTTA	GAATTCACAGGCCAGCTTCC
17	rs3004258	TGAGCGCTTTCCTTCTCAGA	ACAGAGGTCACAGGGTTCAG
20	rs3026703	GGTCTAGTCTCGCTGTGCTA	GCTGTTGTTGACTGTCGGTT
21	rs3026707	ATCCCTCCTGCAGATCTGTG	CAGCCCTCATCTCCAACTGA
22	rs2505989	ATCAACATGGAGGCCTGGAA	AGCTTCTCCTTCCGGCTTAG
24	rs741763	GACAAGAGACAGAGCGCAGA	TTCAGAGAAAGCCCACCTCC
25	rs2505995	ACAGGAAGATGGTGTGAGCC	GGAGAATCTTCAGGGCAGCT
26	rs2506010	CCGCCCTGTCTTTAAACACC	CCATGGCTTCTTCTGTTCCC
28	rs2506020	TGTCCTTCTTGTGGCATAGC	CAATGCCTGTTCAATGCTGT
30	rs2506021	GTAGTGGAGGTGATGGTGCT	AGGTGAATGCTGGTTGGAGA
31	rs2435342	* included with SNP 30 in a single vector	
32	rs2506022	CCTCTACCTCCTGCTGCTTT	CTGCTGGCCAAGAAAACAGG
33	rs2435343	CCTGAGAGTTAAGGGAGGCA	ATTCCCTGTGTCTTCCCTGC
34	rs12247456	GCCTCCCTTGATCCGTGT	TGGTCAGAATTTGGGAAGGG
36	rs7393733	CCATTGTACTGGCTCGTTGA	CTTCCCCATCTCATCACGGT
38	rs2505541	AAATTTCGGCCCTTGTGCTG	ACACTCATTCTCTCAGACGCA

Appendix Table 3: HapMap Samples used as template DNA for each construct.

SNP #	SNP ID	Sample for Non-Risk	Sample for Risk
1	rs788267	NA19445	NA19430
2	rs788263	NA07056	NA06986
3	rs788262	NA19445	NA21118
4	rs788261	NA07056	NA06986
5	rs788260	NA07056	NA06986
6	rs2995411	NA21118	NA19445
7	rs1582227	NA21118	NA19445
8	rs2488278	NA07056	NA06986
10	rs17158318	NA21118	NA19445
11	rs17158320	NA21118	NA19445
12	rs947696	NA21118	NA19445
13	rs7908085	NA20585	NA19445
14	rs10900290	NA19445	NA21118
15	rs947690	NA21118	NA19445
16	rs1547930	NA21118	NA19445
17	rs3004258	NA19445	NA21118
20	rs3026703	NA21118	NA19445
21	rs3026707	NA21118	NA18520
22	rs2505989	NA21118	NA11830
24	rs741763	NA19445	NA11830
25	rs2505995	NA21118	NA19445
26	rs2506010	NA19445	NA11830
28	rs2506020	NA21118	NA19445
30	rs2506021	NA19445	NA11830
31	rs2435342	NA19445	NA11830
32	rs2506022	NA19445	NA11830
33	rs2435343	NA19445	NA11830
34	rs12247456	NA19445	NA19434
36	rs7393733	NA19445	NA19434
38	rs2505541	NA21118	NA19059

6. References

1. Chakravarti A, McCallion AS, Lyonnet S. Hirschsprung disease. In: Beaudet AL, Vogelstein B, Kinzler KW, et al, eds. *The online metabolic and molecular bases of inherited disease*. New York, NY: The McGraw-Hill Companies, Inc.; 2014. http://mhmedical.com/content.aspx?aid=1102916240. Accessed 2016/03/25.

2. Amiel J, Sproat-Emison E, Garcia-Barcelo M, et al. Hirschsprung disease, associated syndromes and genetics: A review. *Journal of Medical Genetics*. 2008;45(1):1-14.

3. Alves MM, Sribudiani Y, Brouwer RWW, et al. Contribution of rare and common variants determine complex diseases—Hirschsprung disease as a model. *Dev Biol*. 2013;382(1):320-329.

4. Badner JA, Sieber WK, Garver KL, Chakravarti A. A genetic study of hirschsprung disease. *Am J Hum Genet*. 1990;46(3):568-580.

5. Sancho E, Batlle E, Clevers H. Signaling pathways in intestinal development and cancer. *Annu Rev Cell Dev Biol*. 2004;20(1):695-723.

6. Emison ES, McCallion AS, Kashuk CS, et al. A common sex-dependent mutation in a RET enhancer underlies hirschsprung disease risk. *Nature*. 2005;434(7035):857-863.

7. Bordeaux M, Forcet C, Granger L, et al. The RET proto-oncogene induces apoptosis: A novel mechanism for hirschsprung disease. *EMBO J*. 2000;19(15):4056-4063. doi: 10.1093/emboj/19.15.4056.

8. Jiang Q, Arnold S, Heanue T, et al. Functional loss of semaphorin 3C and/or semaphorin 3D and their epistatic interaction with ret are critical to hirschsprung disease liability. *The American Journal of Human Genetics*. ;96(4):581-596.

9. Ernst J, Kheradpour P, Mikkelsen TS, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*. 2011;473(7345):43-49.

10. Creyghton MP, Cheng AW, Welstead GG, et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A*. 2010;107(50):21931-21936.

11. Zentner GE, Tesar PJ, Scacheri PC. Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. *Genome Res*. 2011;21(8):1273-1283.

12. Bernstein BE, Stamatoyannopoulos JA, Costello JF, et al. The NIH roadmap epigenomics mapping consortium. *Nat Biotech*. 2010;28(10):1045-1048.

13. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489(7414):57-74.

14. The international HapMap project. *Nature*. 2003;426(6968):789-796.

15. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012;491(7422):56-65.

16. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics*. 2005;21(2):263-265.

17. Purcell S, Neale B, Todd-Brown K, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*. 2007;81(3):559-575.

18. Kapoor A, Jiang Q, Chatterjee S, et al. Population variation in total genetic risk of hirschsprung disease from common RET, SEMA3 and NRG1 susceptibility polymorphisms. *Human Molecular Genetics*. 2015;24(10):2997-3003.

19. Bryne JC, Valen E, Tang ME, et al. JASPAR, the open access database of transcription factor-binding profiles: New content and tools in the 2008 update. *Nucleic Acids Research*. 2008;36(suppl 1):D102-D106.

20. Lang D, Chen F, Milewski R, Li J, Lu MM, Epstein JA. Pax3 is required for enteric ganglia formation and functions with Sox10 to modulate expression of c-ret. *J Clin Invest*. 2000;106(8):963-971.

21. Grice EA, Rochelle ES, Green ED, Chakravarti A, McCallion AS. Evaluation of the RET regulatory landscape reveals the biological relevance of a HSCR-implicated enhancer. *Human Molecular Genetics*. 2005;14(24):3837-3845.

22. Emison ES, Garcia-Barcelo M, Grice EA, et al. Differential contributions of rare and common, coding and noncoding ret mutations to multifactorial hirschsprung disease liability. *The American Journal of Human Genetics*. 2010;87(1):60-74.

23. Welter D, MacArthur J, Morales J, et al. The NHGRI GWAS catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*. 2014;42(D1):D1001-D1006.

24. Pott S, Lieb JD. What are super-enhancers? *Nat Genet*. 2015;47(1):8-12.

25. Andrey G, Montavon T, Mascrez B, et al. A switch between topological domains underlies HoxD genes collinearity in mouse limbs. *Science*. 2013;340(6137).

26. Parveen N, Masood A, Iftikhar N, et al. Comparative genomics using teleost fish helps to systematically identify target gene bodies of functionally defined human enhancers. *BMC Genomics*. 2013;14(1):1-14.

7. Curriculum Vitae

KAMEKO MAKANALANI KARASAKI

Permanent Address: 307 Hamakua Drive Kailua, HI 96734
Email: kkarasa1@jhu.edu
Phone: (808) 542-6873
DOB: October 18th, 1993

EDUCATION

1999- 2012: Punahou School, Honolulu, HI
Graduated with Honors
GPA (unweighted): 3.87

2012-2015: Johns Hopkins University, Baltimore, MD
Bachelor of Science in Molecular and Cellular Biology, Minor in English
GPA: 3.57, BCPM GPA: 3.5, Deans List Awards, Graduated with Honors

2015-present: Johns Hopkins University, Baltimore, MD
Master of Science in Molecular and Cellular Biology
Expected graduation date: May 2016

RESEARCH EXPERIENCE

- September 3, 2013-present: Research Assistant at the **McKusick-Nathans Institute of Genetic Medicine – Dr. Aravinda Chakravarti's Lab**
  - Assisted on the sequencing and functional determination of RET, a gene known to cause Hirschsprung Disease (a congenital disease caused by a lack of neurons in the bowels)
  - Worked on a project to identify potentially causative SNPs in enhancer regions of RET through isolation of ancestral and risk alleles and comparison of enhancer ability through luciferase assays
  - PCR related tasks
    - Primer design using Primer3 and InFusion Primer Design
    - Identifying restriction enzyme cut sites using CLC Main Workbench
  - DNA sequencing
    - Multiple alignments using Sequencher
  - Bacterial transformation and cultures
  - Protein extraction and purification
  - DNA extraction and purification
  - Cell Culture related tasks
    - Luciferase assays using Tecan Infinite F200 Pro
- June 4, 2012-August 1, 2012: Lab Intern at the **University of Hawaii – Manoa Campus – Dr. Yvonne Chan's Lab**
  - Assisted an evolutionary biologist with her research on parrotfish from the Cook Islands
  - PCR related tasks
    - Primer design using Geneious, Primer3, and Primaclade
    - DNA extraction (both destructive and non-destructive on heavily degraded bone samples)
    - DNA sequencing
  - Worked with ancient (therefore heavily degraded) DNA from an archeological dig
  - Learned how to troubleshoot reactions and adjust the concentrations of DNA, buffer, dNTPs, etc, to make amplification more likely
  - Worked extensively in a clean room (with HAZMAT suit) to prevent any external contamination to the degraded samples

MEDICALLY RELATED EXPERIENCE

- June 1, 2015- July 20, 2015: Clinical Shadowing Participant at **Pacific Women's Care LLC – Dr. Charlene Ushijima MD , OB/GYN.**
  - 10 hours/week for 6.5 weeks
  - Observed taking patient histories and physical examinations in office setting
  - Observed active labor and post-partum patient examinations in hospital setting
  - Received tutorial on Implanon and Mirena/ParaGard insertion and removal
  - Observed a variety of procedures, including colposcopy, Implanon insertion

TEACHING EXPERIENCE

- Fall 2015: Biochemisty Lab Teaching Assistant at **The Johns Hopkins University – Dr. Robert Horner**
  - Independently led a three-hour lab for 18 students weekly
  - Provided demos for all experiments performed
  - Responsible for all grading
  - Met with other TAs weekly to practice lab experiments
  - Assisted with proctoring and grading for Biochemistry lecture course
- Spring 2016: General Biology Lab Teaching Assistant at **The Johns Hopkins University – Dr. Rebecca Pearlman**
  - Independently led a three-hour lab for 15 students weekly
  - Responsible for all grading
  - Met with other TAs weekly to discuss lab goals and expectations
  - Assisted with proctoring and grading for General Biology II lecture course

COMMUNITY SERVICE EXPERIENCE

- 2012-2016: Volunteer at **Johns Hopkins University President's Day of Service**
  - Annual day-long volunteer experience to benefit the Baltimore community
  - Projects have ranged from river cleanups to paining elementary school classrooms
  - Learned about some needs of the local community
- 2013-2016: Volunteer at **Art of Caring for Court Appointed Special Advocates (CASA)**
  - Annual afternoon-long silent art auction
  - Have assisted with set up, check in, and bidding
- July 2015: Volunteer at **The Kapiolani Women's Center**
  - Short term volunteer position (24 hours over the course of three weeks)
  - Performed data collection for a best practice study on breast tomosynthesis
  - Assisted with administrative tasks related to an annual Christmas party for breast cancer survivors
- January 2016-Present: Intern at **A Woman's Journey (AWJ) – Johns Hopkins Medicine**
  - 4 hours a week
  - Volunteered at the day-long Baltimore conference in November 2015 as a room docent (attended and monitored four breakout sessions and a lunch plenary session)
  - Assisted with planning and preparation for the eight annual women's health conferences as well as the annual Executive Women's Breakfast
  - Wrote an article on preventive tests to be included in the AWJ program

LEADERSHIP EXPERIENCE

- November 2013-December 2014: **Chief Administrative Officer (CAO)** of the Zeta Chi Chapter of Kappa Alpha Theta Fraternity
  - Responsible for all general mechanics of the chapter
  - In charge of the chapter calendar, correspondence, and other functional necessities
  - Created a weekly update containing information gathered from all other officers through regular communication
  - Planned and executed a workshop (L.E.A.D. - Lead. Empower. Aspire. Develop.) for all officers
  - Oversaw chapter elections and officer transitions
  - Managed chapter roster
  - Attended weekly Executive Board meetings
- Fall 2013-Spring 2015: Senior Event Coordinator for the **JHU Hawai'i 'Ohana Club**
  - Responsible for scheduling all club events and reserving classroom spaces
  - Assisted in planning the annual lu'au, freshman bonding event, and club bonding events
- Fall 2015-Spring 2016: Vice President for the **JHU Hawai'i 'Ohana Club**
  - Worked with other student groups to expand the annual lu'au
  - Discussing the implementation of an educational event to explain the cultural "mixing pot" identity of Hawai'i

SKILLS

- Proficiency in Microsoft Word, Excel, PowerPoint
- Basic knowledge of Adobe Photoshop
- Limited Working Proficiency in Japanese
- Able to work in a Biology wet lab environment
- Excellent communication skills