Supplemental Materials

Supplemental Results

Regions near the *RET* locus identified by MARVEL as significantly associated with S-HSCR

The enhancer-based tests identified 32 significant enhancers, including 23 enhancers near *RET* (within 200kbp from the *RET* TSS). The top S-HSCR-associated enhancer (Chr10:43086011-43087012) contains the well-known HSCR-associated common SNP rs2435357. Another *RET*-locus enhancer (Chr10:43064374-43065375) has VDR binding loss more frequently in cases than in controls (odds ratio: 7.93, 95%CI: 5.88 to 10.70, P<0.0001). VDR is a vitamin-D receptor that has been shown to directly regulate RET expression (Pertile et al. 2018). The promoter-based tests identified only three significant promoters, all of which are promoters of immediate neighboring genes of *RET* (*CSGALNACT2*, *RASGEF1A* and *RP11-351D16.3*). The gene-based tests identified 20 significant genes, including *RET* itself and 15 genes near it (within 700kbp from the *RET* TSS).

Comparing the results of MARVEL with those of existing methods

We compared the enhancer-based results with three commonly used single-variant and region-based association tests using the variants in the hNC enhancers as input. The single-variant Wald test identified 160 variants above the 0.95 confidence interval of the null in the quantile-quantile plot (Supplemental Fig S4a), including 69 variants having an association FDR < 0.1. Comparing the locations of these 160 variants and the 200 loosely-associated enhancers identified by MARVEL, we found 66 of the Wald variants overlapping with 31 of the MARVEL enhancers (Supplemental Fig S4b). Therefore, the Wald test did not identify any S-HSCR-associated variants in the remaining 200-31=169 MARVEL enhancers, including 11 enhancers identified by MARVEL to be significantly associated with S-HSCR. As for the 160-66=94 Wald variants not residing in any of the MARVEL enhancers, 46 of them did not overlap with any sequence motifs. Furthermore, 20 of these 46 variants were in linkage disequilibrium with some loosely associated variants that overlapped sequence motifs, suggesting that they may not be functional themselves. As for the region-based association tests CMC and SKAT-O, none of the hNC enhancers were found to be either significantly or loosely associated (Supplemental Fig S4c-d). These results show that MARVEL can identify S-HSCR associated regions missed by these commonly used methods.

Stage-specific expression of known HSCR genes and genes loosely associated with S-HSCR

We clustered the known HSCR genes and genes loosely associated with S-HSCR based on the mouse trunk NC scRNA-seq data. These genes can be roughly divided into two groups, namely genes with higher expression in neural tube as compared to other stages (mainly in the top cluster in Figure 3C), and genes with higher expression in the migratory stage and autonomic neuron stage (mainly in the bottom cluster). Some genes in these two clusters are involved in NC and HSCR related pathways.

In the top cluster, *Slit1* and *Pax3* are important NC regulators (Szabó and Mayor 2018; Lang and Epstein 2003). *Nkd1* and *Draxin,* both loosely associated with S-HSCR and have similar expression patterns with Pax3, are antagonist in the Wnt-signaling pathway (Ishikawa et al. 2004; Hutchins and Bronner 2018). Furthermore, *Draxin* has been shown to mediate appropriate levels of Wnt signaling for precise regulation of cranial neural crest EMT (Hutchins and Bronner 2018).

The bottom cluster contains two well-known HSCR genes, *Ret and Phox2b*. *Plcg2*, identified in our gene-based analysis, has a similar expression pattern with another known HSCR gene, *Sox10*. The human ortholog of *Plcg2* was previously proposed to be a potential candidate of HSCR (Carrasquillo et al. 2002). In addition to these examples, several other known ENS genes (*Ednrb, Erbb3, Nrp2, Robo1, Sema3d, Slit3*) and genes loosely associated with S-HSCR (*Grb10, Myc, Tcf12)* also have similar expression patterns in the marked stages. Some of these loosely associated genes have been shown to play important roles (Supplemental Table S2) in the pathways in Figure 3A.

Supplemental Discussion

Site-based and region-based methods for studying noncoding genetic variants

Site-based methods (reviewed in Cheng et al. (2020)) aim at prioritizing the variants according to their potential functional effects, based on information such as evolutionary conservation, sequence patterns and epigenomic signals. Such information has been recorded in various annotation databases (Boyle et al. 2012; Ward and Kellis 2016; Watanabe et al. 2017; Khurana et al. 2013), including cell/tissue-specific information indicative of the functional potential of noncoding regions such as chromatin accessibility, histone modifications and transcriptional activities (The ENCODE Project Consortium 2012; The Roadmap Epigenomics Consortium 2015; The FANTOM Consortium and the RIKEN PMI and CLST (DGT) 2014; Andersson et al. 2014). In contrast, region-based methods consider genomic regions of potential functional significance as the basic units, such as enhancers (Luo et al. 2017), promoters (An et al. 2018; Rheinbay et al. 2017), contact regions in the three-dimensional genome architecture (Sallari et al. 2017; Wu and Pan 2019), and combinatorial categories (An et al. 2018). The functional potential of these regions is usually quantified by either the frequency of genetic variants (burden test and its derivatives) or more complex measures, involving sequence kernel association test (Lee et al. 2012), convolutional neural network (CNN) kernels that resemble transcription factor (TF) binding site motifs (Zhou and Troyanskaya 2015; Zhou et al. 2018, 2019), or gene expression models (Gusev et al. 2016; Lou et al. 2019; Zhou et al. 2018).

Cell types for epigenomic profiling

We performed epigenomic profiling using hNCs because HSCR is a kind of neurocristopathies, a disease attributed to the defects in the development of neural crest. We also considered using hNPs, but since hNP cells are quite heterogeneous, containing various intermediates along the neuronal differentiation path (Lau et al. 2019), we focused on the more homogeneous population of hNCs. Additionally, since most noncoding SNPs only have a small effect size, those genetic variants affecting the “earliest” stage of ENS development (i.e. hNC) likely have greater biological impacts than those affecting the late-stage of neuronal differentiation.

Single-cell time-series epigenomic data obtained at different stages of NC development will provide additional information for a more comprehensive evaluation of the functional effects of genetic variants.

Supplemental Methods

Additional details of the MARVEL framework

Required inputs

MARVEL requires two main inputs from the user, namely 1) a list of genetic variants from each subject, which can include single-nucleotide variants and small insertions and deletions, and 2) a set of target regulatory elements. The target regulatory elements required are the enhancers for an enhancer-based analysis, promoters for a promoter-based analysis, and both enhancers and promoters for a gene-based analysis. In the enhancer-based and promoter-based analyses, each regulatory element is considered a target region, while in the gene-based analysis, all the regulatory elements of a gene are considered together as a single virtual target region.

Reconstruction of sample-specific sequences

For each target regulatory region, the genomic sequence of each subject is reconstructed by merging the supplied genetic variants of this subject into the human reference sequence. Specifically, the reconstructed sequence will contain the variant allele if the variant is either homozygous or heterozygous with one of the two alleles being the reference one. For a heterozygous variant with both alleles different from the reference one, one of them is included in the reconstructed sequence arbitrarily.

Motif scores calculation and aggregation

Based on the reconstructed sequence of each target regulatory element, the match (log odds) scores of 771 motifs from the HOCOMOCO (Kulakovskiy et al. 2016) human TF motif database (v11) are computed using MOODS (v1.9.3), with the score set to 0 if the P-value does not pass the default threshold of 10^{-5}. When computing these scores, the nucleotide frequency background is taken from all the sequences in the set of target regulatory elements. If a motif has multiple occurrences in a regulatory element, their match scores are added up to give a single score for this motif. For a gene-based analysis, the scores of a motif in different regulatory elements (including both promoters and enhancers) are further aggregated by a weighted sum, where the weight indicates the strength or confidence of each regulatory element in regulating the gene. For example, if high-throughput chromosome conformation capture data are available, the contact frequencies between the promoter of a gene and different enhancers can be used to define the weights of these enhancers, with a stronger weight given for an enhancer with a higher contact frequency with the promoter.

Validation of MARVEL using simulated data

We simulated 3 motif score profiles to evaluate the performance of GLM-LARS in selecting important motifs caused by different types of genetic variants (Supplemental Fig S1). In each scenario, the data for 100 cases and 100 controls were generated.

The first profile (Supplemental Fig S1a) contained 4 motifs (x1-x4) whose match scores were associated with the phenotype due to common genetic variants with moderate effect sizes and 200 motifs (x5-x204) whose match scores were not affected by the genetic variants (for simplicity, here we use the same symbol for both a motif and its score vector). Among the four associated motifs, x1 and x2 were more associated with the phenotype than x3 and x4, and thus the former two were expected to receive larger absolute coefficients in the regression model.

The second profile (Supplemental Fig S1b) contained 4 motifs (x1-x4) whose match scores were associated with the phenotype due to common variants with moderate effect sizes, 1 motif (x5) whose match scores were associated with the phenotype due to less common variants with large effect sizes, 11 motifs (x6-x16) whose match scores were altered in <1% of individuals due to sequencing errors, and 200 motifs (x17-x216) whose match scores were not affected by the genetic variants. In this scenario, x1 and x2 were expected to have the largest coefficients in the regression model, followed by x5, and in turn followed by x3 and x4.

The third profile (Supplemental Fig S1c) was generated using a procedure similar to the one used for the second profile, except that the last motif (x216) was completely correlated with the first motif (x1), with the match score of the former being half of the match score of the latter in every subject. Both x1 and x216 were expected to receive the same non-zero coefficient in the regression model.

To evaluate the statistical testing procedures of the target regions, we simulated 500 motif score profiles. Among them, 10 motif score profiles were associated with the phenotype using the same procedure as the third profile described above. The remaining 490 motif score profiles were generated randomly with 4 motifs whose match scores were sampled from a normal distribution, 12 motifs whose match scores were altered in <1% of individuals due to sequencing errors, and 200 motifs whose match scores were not affected by variants.

Application of MARVEL to S-HSCR

Whole-genome sequencing and variant calling data

In our previous study (Tang et al. 2018), WGS was performed on 431 S-HSCR cases and 487 ethnically matched controls. Quality checks and processing of the data were performed, and genetic variants were called from the resulting data using standard methods (Tang et al. 2018). We supplied all the identified variants as input to MARVEL, including both common and rare variants.

Production of epigenomic data

ChIP-seq targeting H3K4me1 and H3K27ac and ATAC-seq were performed on the hNC and hPSC with two biological replicates for each assay.

ATAC-seq was performed as previously described (Buenrostro et al. 2015). In brief, around 35,000 FACS-enriched hNC or hPSCs were collected and washed in cold PBS. Transposition reaction was performed according the manufacturer's protocol for Nextra Tn5 transposase Nextra kit (Illumina, Cat. No.: FC-121-1030). Transposed DNA fragments were purified by Qiagen MiniElute PCR purification kit (Qiagen). DNA libraries were then prepared by PCR amplification using NEBNext High-Fidelity PCR kit (New England Biolabs) in the presence of barcoded PCR primers (sequences provided in Buenrostro et al. (2015)). After the PCR amplification, DNA libraries were purified twice by 1.8x AMPure XP beads (Bechman Coulter A63880). The quality of the purified DNA library was assessed by a Bioanalyzer High-Sensitivity DNA Analysis Kit (Agilent). Illumina HiSeq SBS Kit v4 was used for PE101 sequencing (Illumina).

For ChIP-seq, 2.5-5x10^5 FACS enriched hNC or hPSC were collected and fixed with 37% formaldehyde for 10 minutes at room temperature. Chromatin was sonicated by Bioruptor Plus UCD-300 (Diagenode, Belgium). ChIP was performed with 5 µg of H3K4me1 (Abcam, Cat. no.: ab8895) or H3K27ac antibodies (Abcam, Cat. no.: ab4729), and normal IgG (inputs as control), respectively, by MAGnify™ Chromatin Immunoprecipitation System (Invitrogen, USA). ChIP-seq libraries were prepared by MicroPlex Library

Preparation kit v2 (Diagenode) and Illumina sequencing (Pair-End sequencing of 101bp) were done at The University of Hong Kong, Centre for PanorOmic Sciences (HKU, CPOS).

Processing of epigenomic data

The raw data were processed using the ENCODE standard ChIP-seq (https://github.com/ENCODE-DCC/chip-seq-pipeline) and ATAC-seq (https://github.com/kundajelab/atac_dnase_pipelines) pipelines, which included read alignment, quality control, reproducibility assessment, and reads pooling. Narrow peaks were then called from the pooled reads using MACS2 (Zhang et al. 2008) with default settings, involving the matched input controls in the case of ChIP-seq.

Defining target regulatory regions

ChromHMM (Ernst and Kellis 2012) (v1.20) was used to perform genome segmentation based on the ChIP-seq and ATAC-seq peaks. We defined an initial set of enhancers as the genomic segments in chromatin states that emitted both H3K4me1 and H3K27ac marks and overlapped an ATAC-seq peak. These enhancers were then size-normalized to 1kbp each, covering the +/- 500bp regions around the corresponding ATAC-seq peak summits. Target promoters were defined as the +/-500bp regions around all the TSSs of all the genes in GENCODE (v28).

For the negative control study, we collected FANTOM5 (Arner et al. 2015; Andersson et al. 2014) phase 2 permissive enhancers from all samples, extended the length of each enhancer to 1,000 bp while keeping the enhancer center unchanged, and kept only those having no overlap with the active hNC enhancers defined above.

Construction of motif score profiles

For each target enhancer and promoter, we constructed the motif score profile of each subject as described above. For the gene-based analysis, for each gene we considered all its promoters together with all enhancers within 1Mbp from its first TSS. The weight of each regulatory element depends on its genomic distance from the TSS (with the distance of the promoter defined as 0). Specifically, the frequencies of chromatin contact at different distance bins from 0 to 1Mbp were collected from promoter capture Hi-C data produced from hPSCs, obtained from ArrayExpress with accession code E-MTAB-6014 (Montefiori et al. 2018). These frequencies were normalized to have a sum of one and these normalized frequencies were used as the weights (Supplemental Table S4).

Covariates

In the statistical testing procedure of MARVEL, following our previous work (Tang et al. 2018), we used the first three principal components of the genetic variant matrix as the covariates.

Comparing MARVEL with existing single-variant and region-based tests

We compared the enhancer-based results of MARVEL with the results of 3 commonly used association tests, namely the single-variant Wald test and the region-based tests CMC (Combined Multivariate and Collapsing) (Li and Leal 2008) and SKAT-O (Optimized Sequencing Kernel Association) (Lee et al. 2012). All three tests were performed using RVTESTS (Zhan et al. 2016) with the same covariates as MARVEL. Following the common practice (Moutsianas et al. 2015; Lee et al. 2014; Zhang et al. 2019), Wald tests were performed on biallelic common variants with minor allele frequency (MAF) larger than 0.01, while the CMC and SKAT-O tests were conducted on rare variants (MAF<0.01). For both cases, only the variants within the same set of hNC enhancers used by the MARVEL enhancer-based analysis were considered.

The P-values from each testing approach were separately corrected for multiple hypothesis testing using the Benjamini-Hochberg method. Motif scanning was performed on +/-20bp around each variant using MOODS based on the same motif set, P-value threshold and nucleotide background frequencies as in the MARVEL enhancer-based test.

For the loosely associated variants identified by the Wald test, pairwise r^2 values were computed using PLINK (v1.9) with the '--ld' parameter (Chang et al. 2015; Gaunt et al. 2007). The variant pairs with an r^2 value higher than 0.9 were defined to be in linkage disequilibrium.

Analysis of functional pathways

This analysis involved 417 genes loosely associated with S-HSCR from the enhancer-based, promoter-based and gene-based analysis results. For the enhancer-based results, the loosely associated genes included all genes within 50kbp from each loosely associated enhancer, or the gene closest to it when there were none.

The functional interactions were obtained from Reactome (Wu and Haw 2017) (v7.2.3) (https://reactome.org/tools/reactome-fiviz), which contained manually curated functional interactions among over 60% of human proteins. The Reactome-Flviz plugin of Cytoscape (Wu and Haw 2017) was used to obtain and visualize the functional interactions.

To evaluate the functional connectedness of the genes, we counted the number of them having at least one functional interaction with another gene in this set and the total number of interactions among them. We then repeated this same procedure for 1,000 random sets of the same number of genes, and computed the P-value as the number of random sets having a larger number of interactions than the actual genes loosely associated with S-HSCR.

When exploring the functional pathways of these loosely associated genes, we added two sets of genes to the network before looking for highly connected clusters. The first set contained 26 known HSCR genes, including *BACE2, DNMT3B*, *ECE1*, *EDN3*, *EDNRB*, *FAT3*, *GDNF*, *GFRA1*, *KIFBP*, *L1CAM*, *NRG1*, *NRG3*, *NRTN*, *NTF3*, *NTRK3*, *PHOX2B*, *PROK1*, *PROKR1*, *PROKR2*, *PSPN*, *SEMA3A/C/D*, *SOX10*, *TCF4*, and *ZEB2* (Tang et al. 2018; Garcia-Barcelo et al. 2009; Amiel et al. 1996; Tilghman et al. 2019; Alves et al. 2013; Luzón-Toro et al. 2015). The second set contained 19 genes involved in ENS function or NC migration, including *CDC42* (Szabó and Mayor 2018), *CUL1* (Liao et al. 2004), *ERBB2/3/4* (Szabó and Mayor 2018), *GLI3* (Liu et al. 2015), *GNAI1* (Barlow et al. 2003), *GRB2* (Zhang et al. 2011; Alberti et al. 1998), *NRP1/2* (Szabó and Mayor 2018), *PAX3* (Nelms and Labosky 2010; Lang and Epstein 2003), *PLXNB1* (Memic et al. 2018), *ROBO1/2/3* (Szabó and Mayor 2018), *SLIT1/2/3* (Szabó and Mayor 2018), and *TCF12* (Nelms and Labosky 2010). The final network was formed by the genes in these three sets having at least one Reactome functional interaction with each other.

Analysis of mouse trunk NC scRNA-seq data

Processed mouse trunk NC scRNA-seq data (Soldatov et al. 2019) were downloaded from http://pklab.med.harvard.edu/ruslan/neural.crest.html. The expression profiles were extracted from the 'wgm2' data matrix, which had been batch-adjusted and mean-variance normalized as described in the original paper (Soldatov et al. 2019). In-house R (R Core Team 2020) (v3.6.3) scripts and the pheatmap library (https://cran.r-project.org/web/packages/pheatmap/pheatmap.pdf) were used to produce the expression heatmaps. To select genes with stage-specific expression for visualization, principal component analysis was performed on the expression matrix with the genes treated as features. The 10

genes with the largest absolute loading in each of the top three principal components were included in the visualization.

Functional studies

Cell culture

A control hPSC line (UE023A control hPSC line (UE02302) was established as previously described (Lau et al. 2019). hPSCs were maintained in Matrigel (Corning)-coated plate in mTeSR1 medium (Stem Cell Technologies) in a 37 °C humidified 5% CO_2 incubator. The hPSCs were regularly passaged by treating with Dispase (Stem Cell Technologies).

Neural crest induction was performed according to a previously described protocol (Lai et al. 2017). In brief, hPSCs were dissociated into single cell suspension by Accutase (Millipore) and plated on Matrigel-coated plate in a density of 5×10^4 cells cm^{-2} in ES cell medium containing 10 ng/mL fibroblast growth factor 2 (FGF2, Peprotech). The differentiation was started by replacing ES cell medium with KSR medium and gradually switched to N2 medium from day 4 to day 10. To differentiate hPSCs to hNC cells, the cells were treated with 100 nM LDN193189 (Stemgent) from day 0 to day 3, 10 μM SB431542 (Abcam) from day 0 to day 4, 3 μM CHIR99021 (Stemgent) from day 2 to day 10 and 1 μM retinoic acid from day 6 to day 10.

For neuronal differentiation of hNCs to hNPs, hNC cells were dissociated into single cell suspension by Accutase (Millipore) at day 10. For the study of the regulatory element in *PIK3C2B* intron 10, the harvested cells were pelleted and resuspended with N2 medium containing 10 ng/mL FGF2 and 3 μM CHIR99021in a density of 5×10^3 cells μl^{-1}. 2.5 $\times10^4$ hNC cells were seeded as droplets on polyornithine/laminin/fibronectin-coated surface. For the *RET*-associated study, the harvested hNC cells were subjected to fluorescence-activated cell sorting (FACS) and hNC cells which were positive to both HNK-1 (BD Biosciences #560845) and $p75^{NTR}$ (Miltenyi Biotec #130-091-917) were sorted by BD FACSAria III Cell Sorter. 5×10^4 sorted cells were seeded as droplets on polyornithine/laminin/fibronectin-coated surface. Neuronal differentiation was initiated by replacing the medium with N2 medium containing 10 ng/mL BDNF (Peprotech), 10 ng/mL GDNF (Peprotech), 10 ng/mL NT-3 (Peprotech), 10 ng/mL NGF (Peprotech), 1 μM dibutyryl cAMP (Sigma-Aldrich) and 200 μM ascorbic acid (Sigma-Aldrich). The hNC cells differentiated into hNP cells in 9 days.

Plasmid constructions

Human codon-optimized high fidelity Cas9 nuclease construct with GFP tag (pSpCas9(BB)-2A-GFP (PX458)) (Ran et al. 2013) was obtained from Addgene (#48138). Oligos for sgRNA cloning are listed in Supplemental Table S5. The annealed sgRNA oligos were ligated with *Bbs*I-linearized Cas9 construct using Blunt/TA ligation mix (New England Biolabs). For the sgRNA targeting the rs2435357 locus, the annealed sgRNA oligo was ligated with *Afl*II-linearized gRNA cloning vector (Mali et al. 2013) (Addgene #41824) using Blunt/TA ligation mix.

For luciferase assay, *PIK3C2B* intron 10 fragment was amplified from the control hPSC genomic DNA and *NFIA* ORF was amplified from the control hNP cDNA using Q5 Hot Start High-Fidelity DNA Polymerase (New England Biolabs). *PIK3C2B* intron 10 fragment was cloned into NanoLuc luciferase reporter construct (pNL3.2[*NlucP/minP*]) (Promega #N1041) to generate PIK3C2B-pNL construct while *NFIA* ORF was cloned into pFLAG-CMV expression plasmid to generate NFIA-FLAG construct. The *A>T* variant was introduced to PIK3C2B-pNL construct by site-directed mutagenesis using QuickChange Lightning Site-Directed

Mutagenesis Kit (Agilent) to generate PIK3C2B-*A*>*T*-pNL construct. The cloning primers and mutagenesis primers are listed in Supplemental Table S5.

Generation of new hPSC lines using CRISPR-Cas9 system

For the generation of UE-rs2435357 hPSC line, 1×10^6 UE control hPSCs were transfected with 2 μg sgRNA construct, 20 μg ssODNs and 4 μg pSpCas9(BB)-2A-GFP construct using Human Stem Cell Nucleofector Kit 2 (Lonza). For the generation of UE-RASGEF1A-int1-KO and PIK3C2B-int10-KO hPSC lines, 2×10^5 UE control hPSCs were transfected with a pair of pSpCas9(BB)-2A-GFP constructs containing the specific sgRNAs (350 ng per construct) using P3 Primary Cell 4D-Nuclecfector X Kit (Lonza). The transfected cells were plated in Matrigel-coated dish and cultured for 2 days. hPSCs expressing GFP were sorted as single cells into Matrigel-coated 96-well plate with BD FACSAria III Cell Sorter. The sorted cells were expanded for 2 weeks and genotyped to confirm the site-specific conversion or the deletion of the target regions.

Quantitative RT-PCR (RT-qPCR)

Total RNA from hPSCs, hNCs and hNPs was extracted by RNeasy Mini Kit (Qiagen). RNA concentration was determined by NanoDrop 1000 (Thermo Fisher Scientific) and 100ng or 500 ng RNA was then reverse-transcribed to cDNA using HiScript II Q RT SuperMix (Vazyme). The expression levels of the target genes were quantitated using real-time quantitative RT-PCR or Droplet digital PCR (ddPCR). For real-time quantitative RT-PCR, diluted cDNA samples were amplified by Luna Universal Probe qPCR Master Mix (New England Biolabs) using specific TaqMan Gene Expression Assay (*SOX13*, Assay ID: Hs00232193_m1; *ELAVL4*, Assay ID: Hs00956610_mH; *PIK3C2B*, Assay ID: Hs00898499_m1; *PPP1R15B*, Assay ID: Hs03044848_m1; *RET*, Assay ID: Hs01120027_m1; *UBC*, Assay ID: Hs00824723_m1; *18S*, Assay ID: Hs99999901_s1) (Thermo Fisher Scientific) with PCR profiles of 95 °C (1 min) followed by 45 cycles of 95 °C (15 s) and 60 °C (30 s). Fluorescence was measured by ViiA 7 Real-Time PCR System (Thermo Fisher Scientific) at the end of each cycle. Droplet digital PCR (ddPCR) was used to measure *RET* expression in hNPs. 1 μl cDNA samples were mixed with ddPCR Supermix for Probes (Bio-Rad #186-3010) and TaqMan Gene Expression Assay probes of *RET* and *UBC* (Thermo Fisher Scientific). The reaction mixtures were then loaded into the sample wells of DG8 Cartridge (Bio-Rad #186-4008), followed by 70 μl of Droplet Generation Oil for Probes (Bio-Rad #186-3005) into the oil wells. The cartridge was then placed into QX200 Droplet Generator (Bio-Rad) for droplet generation. After droplet generation, the reaction droplets were transferred into a 96-well plate and sealed with foil PCR plate heat seal (Bio-Rad #181-4040) and proceeded to thermal cycling with profiles of 95 °C (10 min) followed by 40 cycles of 95 °C (30 s) and 60 °C (1 min) and then deactivation at 98 °C for 10 min. The end-point fluorescence signals from the reaction droplets were then measured by QX200 Droplet Reader (Bio-Rad). Each individual sample was assayed in triplicate and gene expression was normalized with *UBC* or *18S* expression.

Gel shift assay

3×10^5 HeLa cells were seeded to each well of 6-well plates and cultured in DMEM supplemented with 10% fetal bovine serum and 1% penicillin/streptomycin (Thermo Fisher Scientific) 24 hours before transfection. 2 μg NFIA expression construct (NFIA-FLAG) was transfected to each well of cells by FuGENE HD Transfection Reagent (Promega). Nuclear extracts containing the NFIA protein were extracted from 3 wells of transfected cells using nuclear and cytoplasmic extraction kit (Thermo Fisher Scientific). 1 mM ssODNs (PIK3C2B: 5'- CGC AAG AGC TCT TCA GAA ATG GAT GCC AAG TGT GTC TCC TCT TCC TGA-3' and PIK3C2B-*A*>*T*: 5'- CGC AAG AGC TCT TCA GAA ATG GAT GCC ATG TGT GTC TCC TCT TCC TGA-3') derived from the intron 10 of *PIK3C2B* were biotin-labeled using Biotin 3'-End DNA Labeling Kit (Thermo Scientific).

Biotin-labeled ssODNs were then annealed with reverse complimentary ssODNs to generate biotin-labeled probes. Gel shift assay was performed by mixing the nuclear extracts with biotin-labeled probes according to the manufacture's protocol (LightShift Chemiluminescent EMSA Kit; Thermo Fisher Scientific). In brief, 20 fmol biotin-labeled probes were mixed with 1 µg NFIA nuclear extract in 1X binding buffer containing 50 ng/µl Poly (dI.dC), 0.05% NP-40, 6% glycerol, 60 mM KCl, 1 mM EDTA and 5 mM $MgCl_2$. The binding reactions were incubated for 20 minutes at room temperature. For competition assays, 4 pmol unlabeled probes were added to the mixture before adding the biotin probes. For supershift assays, 0.1 µg anti-NFIA (Sigma-Aldrich, HPA006111) were added to the mixture in the final step before incubation. The binding reactions were resolved in 5% nondenaturing TBE pre-cast gel (Bio-Rad) using Mini-PROTEAN® Electrophoresis System (Bio-Rad) and then transferred to Biodyne B Nylon Membrane (Thermo Fisher Scientific) in 0.5X TBE buffer. After cross-linking, biotin-labeled probes on the membrane were detected using Chemiluminescent Nucleic Acid Detection Module.

Luciferase assay

1.5×10^5 SH-SY5Y cells were seeded to each well of 24-well plates and cultured in 1:1 MEM:F-12 mix supplemented with 10% fetal bovine serum, 1% non-essential amino acids, 1% sodium pyruvate and 1% penicillin/streptomycin (Thermo Fisher Scientific) 24 hours before transfection. 50 ng control firefly luciferase construct (pGL3-control), 150 ng NanoLuc luciferase constructs (pNL3.2 or PIK3C2B-pNL or PIK3C2B-*A*>*T*-pNL) and 25ng NFIA expression construct (NFIA-FLAG) were transfected into the cells using jetPRIME transfection reagent (Polyplus Transfection) according to the manufacturer's protocol. Luciferase activities were detected with Nano-Glo Dual-Luciferase Reporter Assay System (Promega) and measured by VICTOR Nivo Microplate Reader (PerkinElmer).

Supplemental References

Alberti L, Borrello MG, Ghizzoni S, Torriti F, Rizzetti MG, Pierotti MA. 1998. Grb2 binding to the different isoforms of Ret tyrosine kinase. *Oncogene* **17**: 1079–1087.

Alves MM, Sribudiani Y, Brouwer RWW, Amiel J, Antiñolo G, Borrego S, Ceccherini I, Chakravarti A, Fernández RM, Garcia-Barcelo MM, et al. 2013. Contribution of rare and common variants determine complex diseases-Hirschsprung disease as a model. *Dev Biol* **382**: 320–329.

Amiel J, Attié T, Jan D, Pelet A, Edery P, Bidaud C, Lacombe D, Tam P, Simeoni J, Flori E, et al. 1996. Heterozygous endothelin receptor B (EDNRB) mutations in isolated Hirschsprung disease. *Hum Mol Genet* **5**: 355–357.

An J-Y, Lin K, Zhu L, Werling DM, Dong S, Brand H, Wang HZ, Zhao X, Schwartz GB, Collins RL, et al. 2018. Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. *Science* **362**: eaat6576.

Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T, et al. 2014. An atlas of active enhancers across human cell types and tissues. *Nature* **507**: 455–461.

Arner E, Daub CO, Vitting-Seerup K, Andersson R, Lilje B, Drabløs F, Lennartsson A, Rönnerblad M, Hrydziuszko O, Vitezic M, et al. 2015. Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Science* **347**: 1010–1014.

Barlow A, de Graaff E, Pachnis V. 2003. Enteric Nervous System Progenitors Are Coordinately Controlled by the G Protein-Coupled Receptor EDNRB and the Receptor Tyrosine Kinase RET. *Neuron* **40**: 905–916.

Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, Karczewski KJ, Park J, Hitz BC, Weng S, et al. 2012. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res* **22**: 1790–1797.

Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. 2015. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr Protoc Mol Biol* **109**: 21.29.1-21.29.9.

Carrasquillo MM, McCallion AS, Puffenberger EG, Kashuk CS, Nouri N, Chakravarti A. 2002. Genome-wide association study and mouse model identify interaction between RET and EDNRB pathways in Hirschsprung disease. *Nat Genet* **32**: 237–244.

Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**: 7.

Cheng N, Li M, Zhao L, Zhang B, Yang Y, Zheng C-H, Xia J. 2020. Comparison and integration of computational methods for deleterious synonymous mutation prediction. *Brief Bioinform* **21**: 970–981.

Dinsmore CJ, Soriano P. 2018. MAPK and PI3K signaling: At the crossroads of neural crest development. *Dev Biol* **444**: S79–S97.

Ernst J, Kellis M. 2012. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* **9**: 215–216.

Flanagan JG, Vanderhaeghen P. 2002. the Ephrins and Eph Receptors in Neural Development. *Annu Rev*

Neurosci **21**: 309–345.

Garcia-Barcelo M-M, Tang CS, Ngan ES, Lui VC, Chen Y, So M, Leon TY, Miao X, Shum CK, Liu F, et al. 2009. Genome-wide association study identifies NRG1 as a susceptibility locus for Hirschsprung's disease. *Proc Natl Acad Sci* **106**: 2694–2699.

Gaunt TR, Rodríguez S, Day INM. 2007. Cubic exact solutions for the estimation of pairwise haplotype frequencies: implications for linkage disequilibrium analyses and a web tool "CubeX." *BMC Bioinformatics* **8**: 428.

Gunadi, Makhmudi A, Agustriani N, Rochadi. 2016. Effects of SEMA3 polymorphisms in Hirschsprung disease patients. *Pediatr Surg Int* **32**: 1025–1028.

Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BWJH, Jansen R, De Geus EJC, Boomsma DI, Wright FA, et al. 2016. Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet* **48**: 245–252.

Hutchins EJ, Bronner ME. 2018. Draxin acts as a molecular rheostat of canonical Wnt signaling to control cranial neural crest EMT. *J Cell Biol* **217**: 3683–3697.

Ishikawa A, Kitajima S, Takahashi Y, Kokubo H, Kanno J, Inoue T, Saga Y. 2004. Mouse Nkd1, a Wnt antagonist, exhibits oscillatory gene expression in the PSM under the control of Notch signaling. *Mech Dev* **121**: 1443–1453.

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

Kerosuo L, Bronner ME. 2016. cMyc Regulates the Size of the Premigratory Neural Crest Stem Cell Pool. *Cell Rep* **17**: 2648–2659.

Khurana E, Fu Y, Colonna V, Mu XJ, Kang HM, Lappalainen T, Sboner A, Lochovsky L, Chen J, Harmanci A, et al. 2013. Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* **342**: 1235587.

Kubo Y, Baba K, Toriyama M, Minegishi T, Sugiura T, Kozawa S, Ikeda K, Inagaki N. 2015. Shootin1–cortactin interaction mediates signal–force transduction for axon outgrowth. *J Cell Biol* **210**: 663–676.

Kulakovskiy I V., Vorontsov IE, Yevshin IS, Soboleva A V., Kasianov AS, Ashoor H, Ba-Alawi W, Bajic VB, Medvedeva YA, Kolpakov FA, et al. 2016. HOCOMOCO: Expansion and enhancement of the collection of transcription factor binding sites models. *Nucleic Acids Res* **44**: D116–D125.

Lai FP-L, Lau S-T, Wong JK-L, Gui H, Wang RX, Zhou T, Lai WH, Tse H-F, Tam PK-H, Garcia-Barcelo M-M, et al. 2017. Correction of Hirschsprung-Associated Mutations in Human Induced Pluripotent Stem Cells Via Clustered Regularly Interspaced Short Palindromic Repeats/Cas9, Restores Neural Crest Cell Function. *Gastroenterology* **153**: 139-153.e8.

Lang D, Epstein JA. 2003. Sox10 and Pax3 physically interact to mediate activation of a conserved c-RET enhancer. *Hum Mol Genet* **12**: 937–945.

Lau S-T, Li Z, Pui-Ling Lai F, Nga-Chu Lui K, Li P, Munera JO, Pan G, Mahe MM, Hui C-C, Wells JM, et al. 2019. Activation of Hedgehog Signaling Promotes Development of Mouse and Human Enteric

Neural Crest Cells, Based on Single-Cell Transcriptome Analyses. *Gastroenterology* **157**: 1556-1571.e5.

Lee S, Abecasis GR, Boehnke M, Lin X. 2014. Rare-Variant Association Analysis: Study Designs and Statistical Tests. *Am J Hum Genet* **95**: 5–23.

Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, Christiani DC, Wurfel MM, Lin X. 2012. Optimal Unified Approach for Rare-Variant Association Testing with Application to Small-Sample Case-Control Whole-Exome Sequencing Studies. *Am J Hum Genet* **91**: 224–237.

Li B, Leal SM. 2008. Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data. *Am J Hum Genet* **83**: 311–321.

Li N, Kelsh RN, Croucher P, Roehl HH. 2010. Regulation of neural crest cell fate by the retinoic acid and Pparg signalling pathways. *Development* **137**: 389–394.

Liao EH, Hung W, Abrams B, Zhen M. 2004. An SCF-like ubiquitin ligase complex that controls presynaptic differentiation. *Nature* **430**: 345–350.

Liu JAJ, Lai FPL, Gui HS, Sham MH, Tam PKH, Garcia-Barcelo MM, Hui CC, Ngan ESW. 2015. Identification of GLI Mutations in Patients with Hirschsprung Disease That Disrupt Enteric Nervous System Development in Mice. *Gastroenterology* **149**: 1837-1848.e5.

Lou S, Cotter KA, Li T, Liang J, Mohsen H, Liu J, Zhang J, Cohen S, Xu J, Yu H, et al. 2019. GRAM: A GeneRAlized Model to predict the molecular effect of a non-coding variant in a cell-type specific manner ed. Z. He. *PLOS Genet* **15**: e1007860.

Luo Y, de Lange KM, Jostins L, Moutsianas L, Randall J, Kennedy NA, Lamb CA, McCarthy S, Ahmad T, Edwards C, et al. 2017. Exploring the genetic architecture of inflammatory bowel disease by whole-genome sequencing identifies association at ADCY7. *Nat Genet* **49**: 186–192.

Luzón-Toro B, Gui H, Ruiz-Ferrer M, Sze-Man Tang C, Fernández RM, Sham P-C, Torroglosa A, Kwong-Hang Tam P, Espino-Paisán L, Cherny SS, et al. 2015. Exome sequencing reveals a high genetic heterogeneity on familial Hirschsprung disease. *Sci Rep* **5**: 16473.

Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, Norville JE, Church GM. 2013. RNA-Guided Human Genome Engineering via Cas9. *Science* **339**: 823–826.

McLennan R, Krull CE. 2002. Ephrin-As cooperate with EphA4 to promote trunk neural crest migration. *Gene Expr* **10**: 295–305.

Memic F, Knoflach V, Morarach K, Sadler R, Laranjeira C, Hjerling-Leffler J, Sundström E, Pachnis V, Marklund U. 2018. Transcription and Signaling Regulators in Developing Neuronal Subtypes of Mouse and Human Enteric Nervous System. *Gastroenterology* **154**: 624–636.

Montefiori LE, Sobreira DR, Sakabe NJ, Aneas I, Joslin AC, Hansen GT, Bozek G, Moskowitz IP, McNally EM, Nóbrega MA. 2018. A promoter interaction map for cardiovascular disease genetics. *Elife* **7**: e35788.

Moutsianas L, Agarwala V, Fuchsberger C, Flannick J, Rivas MA, Gaulton KJ, Albers PK, McVean G, Boehnke M, Altshuler D, et al. 2015. The Power of Gene-Based Rare Variant Methods to Detect Disease-Associated Variation and Test Hypotheses About Complex Disease ed. S. Ripatti. *PLOS Genet* **11**: e1005165.

Nelms BL, Labosky PA. 2010. Transcriptional Control of Neural Crest Development. *Colloq Ser Dev Biol* **1**: 1–227.

Pertile RAN, Cui X, Hammond L, Eyles DW. 2018. Vitamin D regulation of GDNF/Ret signaling in dopaminergic neurons. *FASEB J* **32**: 819–828.

R Core Team. 2020. R: A language and environment for statistical computing. *R Found Stat Comput , Vienna, Austria*.

Ran FA, Hsu PD, Wright J, Agarwala V, Scott DA, Zhang F. 2013. Genome engineering using the CRISPR-Cas9 system. *Nat Protoc* **8**: 2281–2308.

Rheinbay E, Parasuraman P, Grimsby J, Tiao G, Engreitz JM, Kim J, Lawrence MS, Taylor-Weiner A, Rodriguez-Cuevas S, Rosenberg M, et al. 2017. Recurrent and functional regulatory mutations in breast cancer. *Nature* **547**: 55–60.

Sallari RC, Sinnott-Armstrong NA, French JD, Kron KJ, Ho J, Moore JH, Stambolic V, Edwards SL, Lupien M, Kellis M. 2017. Convergence of dispersed regulatory mutations predicts driver genes in prostate cancer. *bioRxiv* 097451.

Soldatov R, Kaucka M, Kastriti ME, Petersen J, Chontorotzea T, Englmaier L, Akkuratova N, Yang Y, Häring M, Dyachuk V, et al. 2019. Spatiotemporal structure of cell fate decisions in murine neural crest. *Science* **364**: eaas9536.

Szabó A, Mayor R. 2018. Mechanisms of Neural Crest Migration. *Annu Rev Genet* **52**: 43–63.

Tang CS man, Li P, Lai FPL, Fu AX, Lau ST, So MT, Lui KNC, Li Z, Zhuang X, Yu M, et al. 2018. Identification of Genes Associated With Hirschsprung Disease, Based on Whole-Genome Sequence Analysis, and Potential Effects on Enteric Nervous System Development. *Gastroenterology* **155**: 1908-1922.e5.

Tang CSM, Gui H, Kapoor A, Kim JH, Luzón-Toro B, Pelet A, Burzynski G, Lantieri F, So MT, Berrios C, et al. 2016. Trans-ethnic meta-analysis of genome-wide association studies for Hirschsprung disease. *Hum Mol Genet* **25**: 5265–5275.

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57–74.

The FANTOM Consortium and the RIKEN PMI and CLST (DGT). 2014. A promoter-level mammalian expression atlas. *Nature* **507**: 462–70.

The Roadmap Epigenomics Consortium. 2015. Integrative analysis of 111 reference human epigenomes. *Nature* **518**: 317–30.

Tilghman JM, Ling AY, Turner TN, Sosa MX, Krumm N, Chatterjee S, Kapoor A, Coe BP, Nguyen K-DH, Gupta N, et al. 2019. Molecular Genetic Anatomy and Risk Profile of Hirschsprung's Disease. *N Engl J Med* **380**: 1421–1432.

Uribe RA, Hong SS, Bronner ME. 2018. Retinoic acid temporally orchestrates colonization of the gut by vagal neural crest cells. *Dev Biol* **433**: 17–32.

Ward LD, Kellis M. 2016. HaploReg v4: Systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. *Nucleic Acids Res* **44**: D877–D881.

Watanabe K, Taskesen E, van Bochoven A, Posthuma D. 2017. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun* **8**: 1826.

Wu C, Pan W. 2019. Integration of methylation QTL and enhancer-target gene maps with schizophrenia GWAS summary results identifies novel genes. ed. J. Hancock. *Bioinformatics* btz161.

Wu G, Haw R. 2017. Functional interaction network construction and analysis for disease discovery. In *Methods in Molecular Biology*, Vol. 1558 of, pp. 235–253, Humana Press Inc.

Zhan X, Hu Y, Li B, Abecasis GR, Liu DJ. 2016. RVTESTS: an efficient and comprehensive tool for rare variant association analysis using sequence data: Table 1. *Bioinformatics* **32**: 1423–1426.

Zhang J-S, Koenig A, Young C, Billadeau DD. 2011. GRB2 couples RhoU to epidermal growth factor receptor signaling and cell migration ed. C.-H. Heldin. *Mol Biol Cell* **22**: 2119–2130.

Zhang X, Basile AO, Pendergrass SA, Ritchie MD. 2019. Real world scenarios in rare variant association analysis: the impact of imbalance and sample size on the power in silico. *BMC Bioinformatics* **20**: 46.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**: R137.

Zhou J, Park CY, Theesfeld CL, Wong AK, Yuan Y, Scheckel C, Fak JJ, Funk J, Yao K, Tajima Y, et al. 2019. Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk. *Nat Genet* **51**: 973–980.

Zhou J, Theesfeld CL, Yao K, Chen KM, Wong AK, Troyanskaya OG. 2018. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat Genet* **50**: 1171–1179.

Zhou J, Troyanskaya OG. 2015. Predicting effects of noncoding variants with deep learning–based sequence model. *Nat Methods* **12**: 931–934.

Supplemental Tables

Supplemental Table S1 (provided in a separate file) Lists of loosely associated enhancers, promoters and genes identified from the application of MAVEL to the S-HSCR data. Each list contains the locations of the elements in hg38, gene symbols (if applicable), MARVEL association P-values, corresponding FDR Q-values, the TF motifs selected by the procedure, and the AUROC score. For each selected motif, its coefficient in the model is also provided, where a positive value corresponds to an increase of disease risk with a gain of the motif match score, and a negative value corresponds to an increase of disease risk with a loss of the motif match score.

Supplemental Table S2 List of genes relevant to the functional categories identified from the analysis of Reactome functional interactions. The last four columns indicate whether each gene was identified from our enhancer-based (E), promoter-based (P) or gene-based (G) analysis, and whether it passed the FDR<0.1 threshold in at least one of these analyses.

Functional category	Relevant genes loosely associated with S-HSCR	Relevance	E	P	G	FDR< 0.1
Chemotaxis and cell-cell signaling	*RET*	Major HSCR gene; receptor of GNDF; RET signaling regulate enteric NC migration	✓	✓	✓	✓
	FGF3/4/19	FGF-FGFR signaling			✓	
	PLXNB2	Semaphorin-Plexin signaling; Semaphorin has been linked to HSCR (Kapoor et al. 2015; Gunadi et al. 2016; Tang et al. 2016)			✓	
	EPHA2/8	Ephrin signaling (McLennan and Krull 2002; Flanagan and Vanderhaeghen 2002)	✓			
Cell adhesion, migration and integration	*JAM3*	Junctional adhesion protein			✓	
	ACTN1, ACTG1	Subunit of actinin and actin	✓			
	SHTN1	Involved in the CDC42-regulated generation of internal asymmetric signals required for neuronal polarization and neurite outgrowth (Kubo et al. 2015).			✓	
	ITGB7	beta-Integrin; important for cell-ECM interaction	✓			
PI3K/PKC/MAPK signaling	*PIK3C2B*	Subunit of class II PI3K; hub of PI3K pathway (Dinsmore and Soriano 2018)	✓			✓
	MAPK11/12	Hub of MAPK/ERK signaling pathway (Dinsmore and Soriano 2018)			✓	
	PRKCZ	PKC-zeta, subunit of PKC; hub of PKC pathway	✓			
E3 Ubiquitin Ligase complex	*FBXO2/6/15/44*	Subunit of SCF E3 Ubiquitin Ligase (Liao et al. 2004)	✓			✓
Transcriptional regulatory factors	*MYC*	Premigratory NC pool size regulator (Kerosuo and Bronner 2016)	✓			
	ZBTB17	Premigratory NC pool size regulator (Kerosuo and Bronner 2016)	✓		✓	
	RARG	Retinoic acid receptor gamma; Retinoic acid regulates NC migration (Uribe et al. 2018; Li et al. 2010)	✓			

Supplemental Table S3 Recurrent TFs whose motif match scores are significantly more frequently altered in the enhancers loosely associated with S-HSCR than the background of all enhancers. P-values were corrected by the Benjamini-Hochberg method.

TF	FDR Q-value	TF	FDR Q-value	TF	FDR Q-value
ZNF816	0.001285	ZNF768	0.011565	SMAD4	0.0367455
SMAD1	0.001285	ETV3	0.011565	HMGA2	0.0367455
ZNF770	0.001285	ZNF784	0.0148032	MXI1	0.0409594
ZNF219	0.001285	KLF16	0.0156859	GLI1	0.0419602
HAND1	0.0033043	E2F3	0.0156859	ZNF350	0.0419602
PLAGL1	0.0034267	RARG	0.0156859	JUNB	0.0419602
GLI2	0.0034267	SP4	0.02056	ZNF274	0.0419602
ELK4	0.0056073	CTCF	0.02056	E2F2	0.0448582
NHLH1	0.0056073	TEAD2	0.02056	ZNF140	0.0448582
POU5F1	0.0068533	ZNF136	0.02056	CREB3	0.0448582
SP1	0.0068533	TEAD4	0.0220286	ZNF467	0.04626
TFAP2B	0.0068533	SP3	0.0227017	GLI3	0.04626
E2F4	0.0068533	THAP1	0.0250054	ZIC2	0.04626
ETV1	0.0068533	VEZF1	0.0257676	ZNF354A	0.04626
ZNF148	0.00771	SMAD2	0.0281415	THRA	0.04626
ZBTB17	0.0088114	SP2	0.0281415	ZSCAN22	0.04626
SMAD3	0.0088114	ZBED1	0.0287715	HOXC6	0.0478261
SP1	0.011565	ZNF317	0.034438	MECOM	0.0480392

Supplemental Table S4 Weights of regulatory elements used in the gene-based analysis of the S-HSCR study with respect to their distance from the gene TSS.

Bin (distance from TSS)	Weight	Bin (distance from TSS)	Weight
0kb - 50kb	0.132258	500kbp - 550kbp	0.022819
50kbp - 100kbp	0.133288	550kbp - 600kbp	0.017064
100kbp - 150kbp	0.135888	600kbp - 650kbp	0.013075
150kbp - 200kbp	0.126787	650kbp - 700kbp	0.010261
200kbp - 250kbp	0.107039	700kbp - 750kbp	0.007829
250kbp - 300kbp	0.085864	750kbp - 800kbp	0.006324

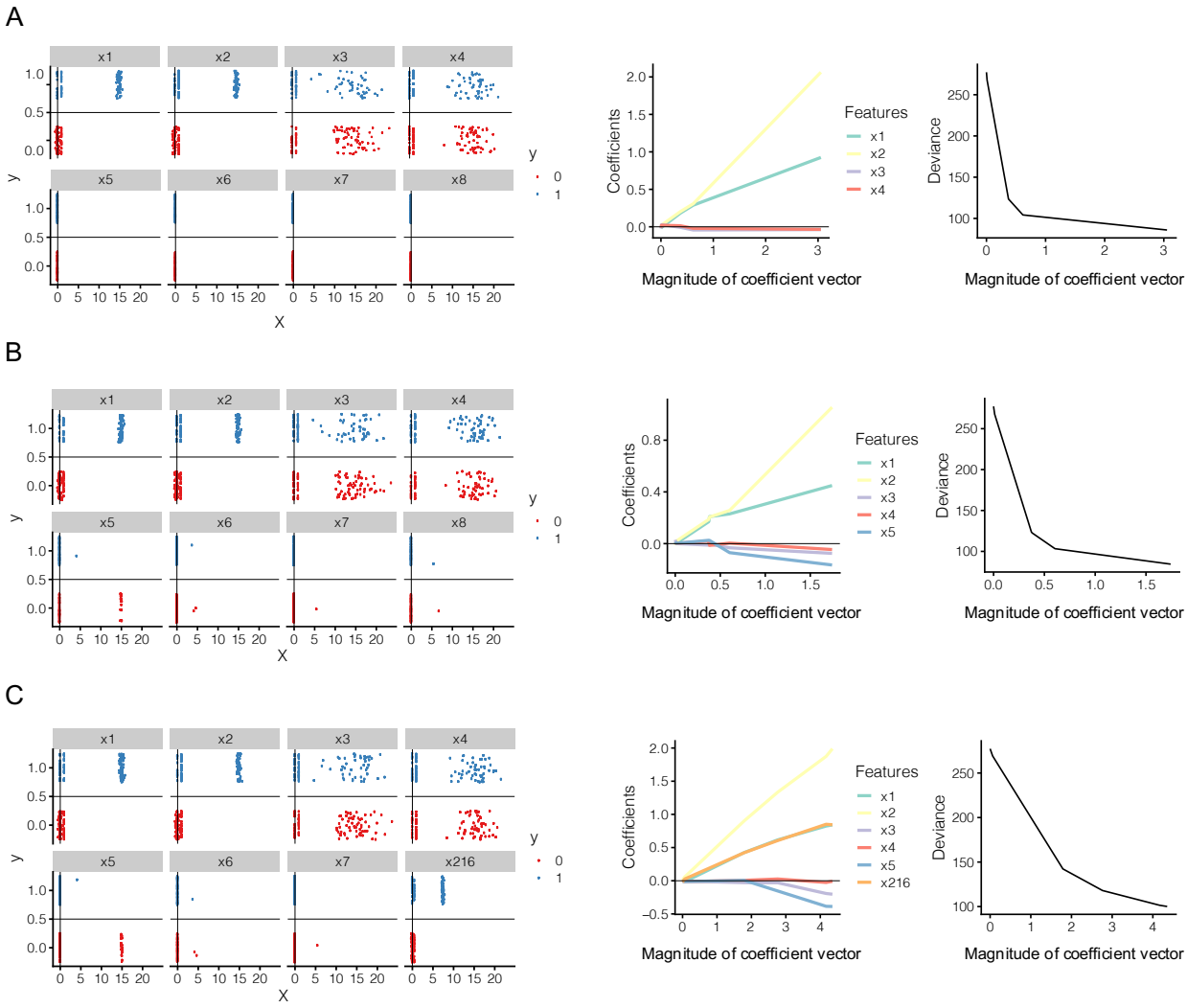
300kbp	- 350kbp	0.066000	800kbp	- 850kbp	0.004990
350kbp	- 400kbp	0.051969	850kbp	- 900kbp	0.004070
400kbp	- 450kbp	0.038727	900kbp	- 950kbp	0.003359
450kbp	- 500kbp	0.029764	950kbp	- 1000kbp	0.002626

Supplemental Table S5 List of primers and oligos used for cloning, genotyping and mutagenesis. *: Sequences in small letters represent the complementary overhangs or restriction enzyme sites for the purpose of cloning.

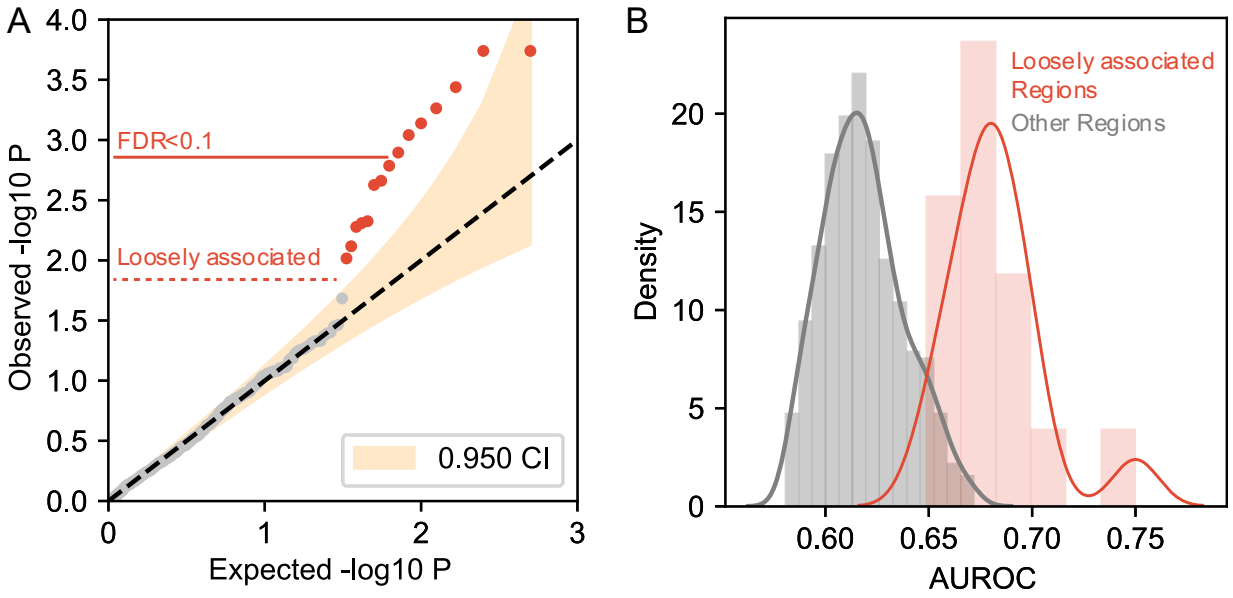
Primer name	**Sequence (5' to 3')***	**Purpose**
PIK3C2B-int10-KO-5'gRNA-F	caccgCACCTACACACTCTAGCAAC	For cloning of 5' sgRNA for deletion of *PIK3C2B* intron 10
PIK3C2B-int10-KO-5'gRNA-R	aaacGTTGCTAGAGTGTGTAGGTGc	For cloning of 5' sgRNA for deletion of *PIK3C2B* intron 10
PIK3C2B-int10-KO-3'gRNA-F	caccgTTGAGAGCCCAGGCCAGTAA	For cloning of 3' sgRNA for deletion of *PIK3C2B* intron 10
PIK3C2B-int10-KO-3'gRNA-R	aaacTTACTGGCCTGGGCTCTCAAc	For cloning of 3' sgRNA for deletion of *PIK3C2B* intron 10
PIK3C2B-int10-KO-seq-F	CTTCCCACTGAAGGCTGACA	For genotyping of the deletion of *PIK3C2B* intron 10 and RT-PCR of *PIK3C2B*
PIK3C2B-int10-KO-seq-R	AGAGCAGTTCCCTTACCTCG	For genotyping of the deletion of *PIK3C2B* intron 10
PIK3C2B-exon12-R	AGGGCTTCCACGACCTTCT	For RT-PCR of *PIK3C2B*
NFIA-FL-FLAG-F	TTGgaattcaATGTATTCTCCGCTCTGTCTCACC	For cloning of NFIA expression construct
NFIA-FL-FLAG-R	TTTgtcgacTTATCCCAGGTACCAGGACTGTG	For cloning of NFIA expression construct
PIK3C2B-int10-pNL-F	TTTggtaccGAAGAGAGGGATACTGTCAGGTAAC	For cloning of *PIK3C2B* intron 10 to pNL
PIK3C2B-int10-pNL-R	TTTctcgagAGCCAAGACACTGGGATCCTG	For cloning of *PIK3C2B* intron 10 to pNL
PIK3C2B-*A*>*T*-pNL-F	AGGAAGAGGAGACACACATGGCATCCATTTCTGAA	For mutagenesis of PIK3C2B-pNL
PIK3C2B-*A*>*T*-pNL-R	TTCAGAAATGGATGCCATGTGTGTCTCCTCTTCCT	For mutagenesis of PIK3C2B-pNL
RET-r2435357-*T*-allele-gRNA-F	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCCTGTGGATGACCATGTAA	For cloning of sgRNA for conversion of rs2435357 from C>T
RET-r2435357-*T*-allele-gRNA-R	GACTAGCCTTATTTTAACTTGCTATTTCTAGCTCTAAAACTTACATGGTCATCCACAGGC	For cloning of sgRNA for conversion of rs2435357 from C>T

RET_rs2435357_*C>T*_ssODN	AGGCCTGGCTGAACAGGACTGGCCACCCAAGTGGCCTGTGGATGACCATGTAAGGGTCACTGGCCCCCTTGGCTGCAGGGCTGTAGAGTCTGCCCCAGCT	Single oligonucleotides for conversion of rs2435357 from C>T
RASGEF1A-int10-KO-5'gRNA-F	caccgTGGACTCCTGCCGGCAACCA	For cloning of 5’ sgRNA for deletion of *RASGEF1A* intron 1
RASGEF1A-int1-KO-5'gRNA-R	aaacTGGTTGCCGGCAGGAGTCCAc	For cloning of 5’ sgRNA for deletion of *RASGEF1A* intron 1
RASGEF1A-int1-KO-3'gRNA-F	caccgCGACCCAGCCTCTGACCTAC	For cloning of 3’ sgRNA for deletion of *RASGEF1A* intron 1
RASGEF1A-int1-KO-3'gRNA-R	aaacGTAGGTCAGAGGCTGGGTCGc	For cloning of 3’ sgRNA for deletion of *RASGEF1A* intron 1
RET-int1-seq-F	CAGGGCCAGTGAACAATGTA	For genotyping of rs2435357
RET-int1-seq-R	ACCACCCACACTTCCATACC	For genotyping of rs2435357
RASGEF1A-int1-seq-F	GTTGACCTGGGGAGAGATGT	For genotyping of the deletion of *RASGEF1A* intron 1
RASGEF1A-int1-seq-R	AAGAATCTTTCCCCGCTGCA	For genotyping of the deletion of *RASGEF1A* intron 1

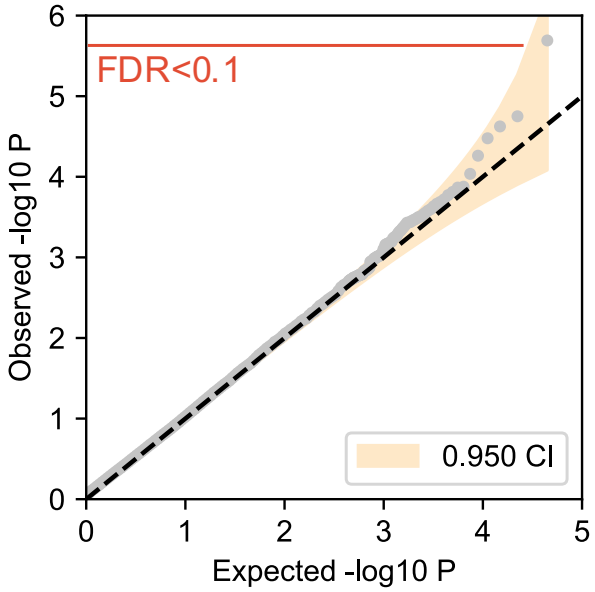
Supplemental Figures



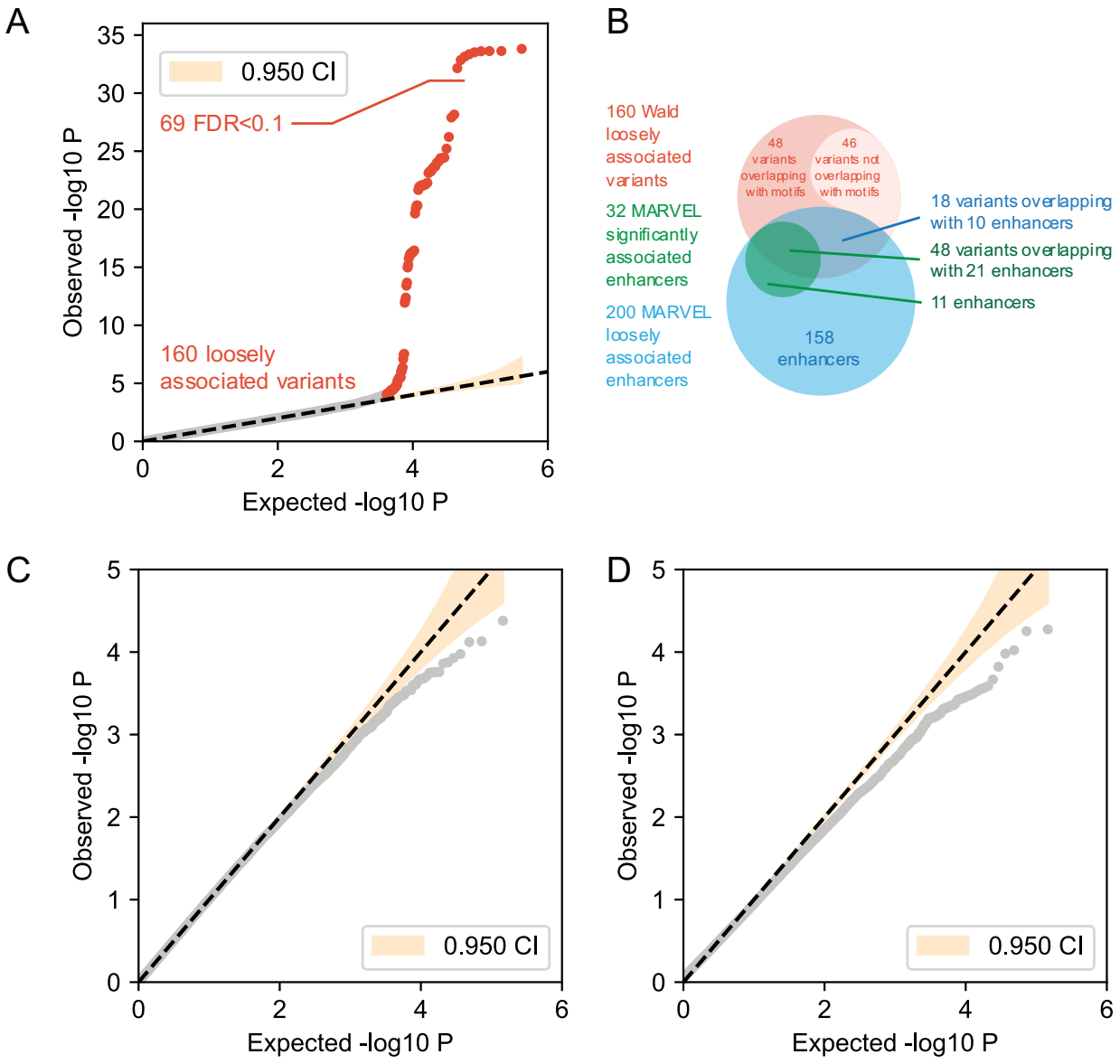
Supplemental Fig S1 Simulated motif score profiles for testing the feature selection procedure of MARVEL. (**A**-**C**) The three motif score profiles reflect situations in which the match scores of some motifs are affected by phenotype-associated independent common variants with moderate effect sizes (A), independent common variants with moderate effect sizes and less common variants with large effect sizes (B), and variants that are correlated (C). Details of the simulation procedure are given in Methods. In each panel, the left scatterplots show the match scores of the motifs affected by phenotype-associated variants and some of the motifs that are not. Random jitters are added to reduce overlapping of points. The phenotype-associated motifs were designed to be x1-x4 (A), x1-x5 (B), and x1-x5 and x216 (C). The middle graphs show the coefficient values of different motifs when the magnitude of the coefficient vector is set to different values. Only motifs with non-zero coefficients are shown. The right graphs show the change of deviance at different magnitude values of the coefficient vector.



Supplemental Fig S2 Verification of the statistical testing procedures of MARVEL based on simulated data. (**A**) Quantile-quantile plot of simulated motif score profiles. Motif scores were generated for 500 regulatory regions, each with 216 motifs, 100 cases and 100 controls. For 10 of these regions, their motif score profiles were generated using the same way for generating the third motif score profile for testing the feature selection procedure of MARVEL. For the remaining 490 regions, their motif score profiles were generated randomly. Each dot in the plot corresponds to one regulatory region. The yellow shaded area shows the 95% confidence interval according to beta error distribution. (**B**) Comparison of the AUROC value distributions of the loosely associated regions identified by MARVEL with the other regions.



Supplemental Fig S3 Quantile-quantile plot of the association P-values of the background set of enhancers.



Supplemental Fig S4 Comparing the results of MARVEL and three commonly used association tests when applied to the hNC enhancer variants. (**A**) Q-Q plot of the association P-values produced by the single-variant Wald test. (**B**) Venn diagram showing the overlap between the variants identified by the Wald test and the enhancers identified by MARVEL. (**C**-**D**) Q-Q plots of the association P-values produced by the region-based tests CMC (C) and SKAT-O (D).