Published in final edited form as:
*Nat Genet*. 2024 May ; 56(5): 877–888. doi:10.1038/s41588-024-01717-7.

# STR mutations on chromosome 15q cause thyrotropin resistance by activating a primate-specific enhancer of *MIR7–2/ MIR1179*

**Helmut Grasberger**[1], **Alexandra M. Dumitrescu**[2,3], **Xiao-Hui Liao**[2], **Elliott G Swanson**[4,5], **Roy E. Weiss**[6], **Panudda Srichomkwun**[2], **Theodora Pappa**[2], **Junfeng Chen**[7], **Takashi Yoshimura**[7], **Phillip Hoffmann**[8], **Monica Malheiros França**[2], **Rebecca Tagett**[9], **Kazumichi Onigata**[2], **Sabine Costagliola**[10], **Jane Ranchalis**[4], **Mitchell R Vollger**[4], **Andrew B Stergachis**[4,5,11], **Jessica X. Chong**[4,11], **Michael J Bamshad**[4,5,11], **Guillaume Smits**[8,12], **Gilbert Vassart**[10], **Samuel Refetoff**[2,13,14,✉]

[1]Department of Internal Medicine, Medical School, University of Michigan, Ann Arbor, Ml, USA.

[2]Department of Medicine, The University of Chicago, Chicago, IL, USA.

[3]Committee on Molecular Metabolism and Nutrition, The University of Chicago, Chicago, IL, USA.

[4]Division of Medical Genetics, Department of Medicine, University of Washington, Seattle, WA, USA.

[5]Department of Genome Sciences, University of Washington, Seattle, WA, USA.

[6]Department of Medicine, University of Miami Miller School of Medicine, Miami, FL, USA.

[7]Institute of Transformative Bio-Molecules (WPI-ITbM) and Graduate School of Bioagricultural Sciences, Nagoya University, Nagoya, Japan.

[8]Interuniversity Institute of Bioinformatics in Brussels, Universite Libre de Bruxelles-Vrije Universiteit Brussel, Brussels, Belgium.

[9]Michigan Medicine BRCF Bioinformatics Core, University of Michigan, Ann Arbor, Ml, USA.

[10]Institut de Recherche Interdisciplinaire en Biologie Humaine et Moléculaire (IRIBHM), Universite Libre de Bruxelles, Brussels, Belgium.

**Reprints and permissions information** is available at www.nature.com/reprints.

✉ **Correspondence and requests for materials** should be addressed to Samuel Refetoff., refetoff@uchicago.edu.

Author contributions
H.G. and S.R. conceptualized the study, organized the data, prepared the figures and cowrote the original draft of the paper. S.R. and G.V. identified the first families, and advised and helped in devising all studies. S.R., A.M.D. and R.E.W. recruited, examined and followed participants and family members. G.S. analyzed the initial WGS data and identified the STR variant in humans with RTSH. He was assisted by P.H. H.G. and X.-H.L. designed and performed the laboratory studies. H.G. and R.T. analyzed the RNA-seq data. P.S., T.P., M.M.F. and K.O. performed gene sequencing and analyzed the results. J.C. and T.Y. obtained tissues, and sequenced and analyzed the *Gorilla* DNA samples. J.X.C. and M.J.B. performed WGS, analyzed the results and wrote parts of the paper. J.R. and A.B.S. performed the Fiber-seq reaction. M.R.V., E.G.S. and A.B.S. performed the Fiber-seq data analysis. S.C. advised and planned some of the studies. All authors reviewed and edited the paper.

Competing interests
A.B.S. is a co-inventor on a patent relating to the Fiber-seq method. The other authors declare no competing interests.

**Extended data** is available for this paper at https://doi.org/10.1038/s41588-024-01717-7.

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41588-024-01717-7.

[11]Brotman-Baty Institute for Precision Medicine, Seattle, WA, USA.

[12]Center of Human Genetics, Hôpital Erasme, Hôpital Universitaire de Bruxelles, and Department of Genetics, Hôpital Universitaire des Enfants Reine Fabiola, Hôpital Universitaire de Bruxelles, Université Libre de Bruxelles, Brussels, Belgium.

[13]Committee on Genetics, The University of Chicago, Chicago, IL, USA.

[14]Department of Pediatrics, The University of Chicago, Chicago, IL, USA.

## Abstract

Thyrotropin (TSH) is the master regulator of thyroid gland growth and function. Resistance to TSH (RTSH) describes conditions with reduced sensitivity to TSH. Dominantly inherited RTSH has been linked to a locus on chromosome 15q, but its genetic basis has remained elusive. Here we show that non-coding mutations in a $(TTTG)_4$ short tandem repeat (STR) underlie dominantly inherited RTSH in all 82 affected participants from 12 unrelated families. The STR is contained in a primate-specific/*Alu* retrotransposon with thyroid-specific *cis*-regulatory chromatin features. Fiber-seq and RNA-seq studies revealed that the mutant STR activates a thyroid-specific enhancer cluster, leading to haplotype-specific upregulation of the bicistronic *MIR7–2/MIR1179* locus 35 kb downstream and overexpression of its microRNA products in the participants' thyrocytes. An imbalance in signaling pathways targeted by these micro-RNAs provides a working model for this cause of RTSH. This finding broadens our current knowledge of genetic defects altering pituitary-thyroid feedback regulation.

Thyroid hormone is critical fornormal fetal and postnatal development, and thyroid function is controlled by a pituitary-thyroid feedback loop. Routine measurement of pituitary-derived thyrotropin (TSH) in blood is widely used in neonatal screening for the detection of the most common inborn endocrine disorder, congenital hypothyroidism[1]. TSH elevation (hyperthyrotropinemia) correlates in most instances with low thyroid hormone, which can result in impaired mental and growth development with irreversible consequences if not diagnosed and treated early[2]. Consequently, hyperthyrotropinemia at birth triggers the initiation of thyroid hormone treatment and often results in its life-long maintenance. However, conditions within the spectrum of resistance to TSH (RTSH; MIM 275200) have hyperthyrotropinemia associated with normal (NL) thyroid hormone levels as compensated hypothyroidism[3,4]. RTSH typically manifests with NL-sized or hypoplastic thyroid gland in situ, in contrast to thyroid hormone synthesis defects that manifest with TSH-induced thyroid gland enlargement.

In 1968, RTSH was recognized by Stanbury and Rocmans[5]. Molecular cloning of the TSH receptor in 1989 (ref. 6) allowed the demonstration of compensated RTSH caused by compound heterozygous *TSHR* mutations[7]. Dominantly inherited RTSH without *TSHR* mutations was first identified in our laboratory[8,9]. Using genetic linkage in five families, a 2.9-Mb RTSH locus was mapped on chromosome (chr) 15q25.3–26.1 (refs. 10,11). We now report that this dominantly inherited RTSH (MIM 609893) is caused by mutations within a primate-specific *cis*-regulatory element at this locus. Using Fiber-seq on thyroid tissue, we were able to demonstrate the activation of codependent enhancer elements that selectively upregulate a neighboring microRNA locus in the affected individuals.

## Results

### Mutations in an *Alu*-derived STR cause RTSH

We performed whole-genome sequencing (WGS) in the five original RTSH families with significant linkage to the locus on chrl5q (Fig. 1a). Because no genes were identified in which multiple families had protein-coding variants segregating with RTSH, we screened for potential mutations by mapping rare non-coding variants segregating with RTSH in individual families. Only one rare variant was shared by more than two families. This variant, an ultrarare deletion (Table 1) in a non-coding short tandem repeat (STR) located between *DET1* and *MIR7–2,* was found to segregate with the RTSH phenotype in four of the five families (24, 25, 26 and 35; Fig. 2). The wild-type (WT) STR (STR$^{wt}$) allele consists of four tandem base pairs (TTTG)$_4$ repeats. The unaffected individuals in each family were homozygous for four repeats (STR 4/4), whereas all affected individuals were heterozygous for a deletion of one repeat (STR 4/3; Fig. 1b). In the fifth family screened by WGS (30; Fig. 2), RTSH segregated with a T > G substitution (absent in gnomAD; Extended Data Table 1) in the third repeat; TTGG instead of TTTG (further referred to as STR 4/4*; Fig. 1b and Table 1).

Subsequently, other families with potentially inherited RTSH were screened for STR variants by either WGS or Sanger sequencing. Seven additional families were found to have STR variants, six with STR3 and one with STR4* (Fig. 2). The finding of pathogenic STR variants in 12 of 101 families with RTSH makes this STR the second most common locus underlying RTSH after *TSHR* (Fig. 3a). In the 12 families, homozygosity for the STR$^{wt}$ was confirmed in all 66 unaffected relatives tested, and heterozygosity for either STR3 or STR4* was confirmed in all 76 affected participants as well as 6 individuals whose thyroid phenotype could not be assessed because they were taking levothyroxine (L-T4) and no pretreatment information was available. The lack of recent shared ancestry among families with STR3 (Supplementary Table 1) and the demonstration that the deletion of one repeat arose de novo in family 14 (Fig. 2) indicate that STR3 is a recurrent variant. The observation of the same phenotype in individuals with two different ultrarare STR variants, STR3 and STR4* (referred to from here on as STR$^{mut}$) in 12 of 101 families with RTSH is unlikely to be due to chance (false discovery rate (FDR) = $2.8 \times 10^{44}$ based on the frequency of STR4/3 in gnomAD; Table 1).

Although the STR falls within a 1-kb region that is among the 10% most-constrained regions in terms of variation in the human genome (Table 1), it is not conserved in vertebrates (mean 100 vertebrates PhyloP = −0.146) due to its location in a primate-specific *Alu*Sxl repeat element that is present only in old-world monkeys and apes (Extended Data Fig. 1).

Interestingly, gorillas have significantly higher serum TSH levels than other great apes while maintaining NL thyroid hormone levels, that is, an RTSH-like phenotype[12]. Sequencing the equivalent STR in 11 gDNA samples of gorillas in captivity revealed 1 was homozygous for (TTTG)$_4$ corresponding to the human WT allele, 7 homozygous for TCTG in the fourth repeat (4 of these were related) and 3 heterozygous TTTG/TCTG (Fig. 1c). Unfortunately, no paired serum samples were available for genotype-phenotype correlation. Nevertheless, identification that seven of eight unrelated gorillas were heterozygous or homozygous for a variant STR and that the *Gorilla* population seems to exhibit an RTSH-like phenotype supports a conserved function of the STR in other primates.

The thyroid phenotypes of participants with STR$^{mut}$ were assessed in untreated and >6 weeks off L-T4 individuals from the 12 families. All participants with STR$^{mut}$ had serum TSH values above the upper limit of the reference range indicating complete penetrance (Fig. 3b). Average serum free thyroxine (FT4) was slightly lower in participants with STR$^{mut}$ but, except for one individual, still within the population reference range (Fig. 3b and Supplementary Fig. 1). There was no correlation between TSH and FT4 values in participants with STR$^{mut}$ (Fig. 3c). TSH and FT4 were highest in newborns with STR$^{mut}$ and showed the expected age-related reduction (Fig. 3e). Consistent with RTSH, thyroid gland size was reported as normal or small on physical examination and confirmed by ultrasound (5 being normal in size and 4 hypoplastic). Radionuclide uptake was obtained in seven participants with STR$^{mut}$ and found to be low in six and normal in one. Yet, two unique features set participants with STR$^{mut}$ apart from *TSHR* mutation carriers. First, most participants with STR$^{mut}$ had high serum thyroglobulin (TG) levels that were uncorrelated to the degree of hyperthyrotropinemia (Fig. 3b,d). Second, a subset of STR$^{mut}$ carriers appeared to be paradoxically prone to develop symptomatic proliferative thyroid disease later in life. In fact, 3 of 82 STR$^{mut}$ heterozygous individuals (3.66%) required surgery because of enlarged nodular glands. For comparison, thyroidectomy for non-toxic nodular goiters and benign thyroid adenomas is performed at a rate of 0.016% per year in the general US population[13]. Thyroid tissue was obtained from two of them for further analysis, participants 11–4 from family 60 heterozygous for STR3 and 1–2 from family 30 heterozygous for STR4* (Figs. 2 and 3f).

## STR$^{mut}$ activates a thyroid-specific enhancer

The STR maps to a predicted *cis*-regulatory element (SCREEN database; EH38E3153006) that exhibits a weak thyroid-specific enhancer-like signature and is located within a topologically associating domain (TAD) containing seven annotated genes, several of them being normally expressed in thyroid tissue (Extended Data Fig. 2 and Supplementary Fig. 2).

To directly test whether STR$^{mut}$ is associated with alterations in the activity of this *cis*-regulatory element, we performed single-molecule chromatin fiber sequencing (Fiber-seq)[14–16] on thyroid tissue from two individuals homozygous for the STR$^{wt}$ allele, a healthy control and one with multinodular goiter (MNG), as well as the two RTSH individuals heterozygous for STR$^{mut}$ as described above (Fig. 4 and Extended Data Figs. 3 and 4).

Fiber-seq chromatin accessibility patterns from the normal control individual and the STR$^{wt}$ haplotype from the participant with RTSH mirrored previously published DNase-seq data from NL thyroid tissue (Fig. 4b) and confirmed a weak site of chromatin accessibility directly overlapping the STR (Fig. 4c). In contrast, the STR$^{mut}$ haplotype was associated with a strong peak of chromatin accessibility directly overlapping the STR$^{mut}$ site, as well as four additional accessible chromatin elements clustered within 6 kb surrounding theSTR$^{mut}$ location (Fig. 4c and Extended Data Fig. 3).

To evaluate whether this regulatory element cluster enhances the accessibility of any gene promoters along chrl5, we generated short-read WGS data from both parents of this individual (1–1 and 1–2 from family 60) to enable the haplotype phasing of Fiber-seq data across the entire chrl5 (ref. 17). This demonstrated that chromatin accessibility at a putative *MIR7–2/MIRU79* promoter located ~35 kb downstream of STR$^{mut}$ was substantially and selectively enhanced on the STR3 haplotype (Fig. 4c,d), with other promoters within the same TAD showing unchanged chromatin accessibility between the STR$^{mut}$ and STR$^{wt}$ haplotypes. The STR4* haplotype is associated with the same chromatin accessibility pattern as STR3 (Extended Data Fig. 3c).

## STR$^{mut}$ creates an active enhancer cluster

To determine how the STR$^{mut}$ haplotype creates this enhancer cluster, we analyzed the codependency between the single-molecule accessibility of each enhancer cluster element. Overall, we found that the accessibility of these five elements is strongly codependent (Fig. 4e), consistent with this cluster being co-actuated selectively along the STR$^{mut}$ haplotype.

To evaluate the importance of each of the five elements for the accessibility of the codependent enhancer cluster, we simulated epigenetic editing for each element, focusingon identifying fibers where each element was in a closed state. Then, we analyzed the dependency among the remaining four elements in these closed-state fibers (Fig. 4f). Notably, when we silenced the regulatory element overlapping with STR$^{mut}$ (element 2), it substantially disrupted the accessibility of the enhancer cluster, unlike silencing the other four elements, suggesting that element 2 has a central role in activating this enhancer cluster.

Individual chromatin fibers harboring actuated element 2 similarly had hypo-CpG methylation of the surrounding CpG dinucleotides (Extended Data Fig. 4). Of note, skin fibroblasts from individuals with RTSH have hyper-CpG methylation at these elements, suggesting that a thyroid-specific regulatory protein is likely contributing to the formation of hypo-CpG methylated DNA and accessible chromatin at this enhancer cluster (Extended Data Fig. 5).

To explore how the STR$^{mut}$ variant increases chromatin accessibility of its overlying regulatory element, we evaluated this regulatory element for transcription factor (TF) footprints. We found that the STR3 variant was contained within a 30–60 bp footprint on 22 of 27 chromatin fibers overlapping that site (Fig. 1a). We also observed that 4 of 16 chromatin fibers from the STR$^{wt}$ haplotype contained a similarly sized footprint at that site. Notably, the STR4* haplotype was similarly associated with the activation of an overlying accessible chromatin patch and the stabilization of a 30–60 bp footprint (Fig. 1c). Thus, the STR$^{mut}$ variant turns what is normally an infrequently bound TF-binding element along the STR$^{wt}$ haplotype into one that is strongly bound (Fig. 1b).

Although the specific TFs that cause this footprint are unknown, the 30–60 bp footprint contains adjacent predicted binding elements for the thyroid-specific TF FOXE1 (refs.

18,19; Fig. 1d). While the STR$^{wt}$ haplotype similarly contains potential FOXE1 binding elements, the spacing between these elements is altered along the STR$^{mut}$ haplotype. Similarly, the STR4* and the *Gorilla* (TTTG)$_3$(TCTG) sequences alter the spacing between these FOXE1 binding elements (Fig. 1d). As the winged-helix DNA binding domain of FOXE1 permits TF occupancy of DNA wrapped around a nucleosome[20], alterations to the spacing of these adjacent elements could markedly impact the pioneering activity of FOXEl at this site by changing the helical pitch between cobound FOXE1 elements.

### The STR$^{mut}$-activated enhancer induces a bicistronic miRNA locus

We performed RNA-seq in search for transcripts that are differentially expressed in STR$^{mut}$ samples versus STR$^{wt}$ controls. Non-adenomatous thyroid tissuesamples from two participants with STR$^{mut}$ were analyzed by standard and small RNA-seq. STR$^{wt}$ controls comprised five NL thyroids (biopsies obtained during surgery for parathyroid adenomas) and one MNG thyroid.

We found that protein-coding genes within the RTSH locus had low expression in the thyroid and were not differentially expressed in the participants' thyroid glands (Supplementary Fig. 3). However, a 13.5-kb region of interest (ROD that is 35 kb q-terminal of the STR and contains the miRNA promoter with increased chromatin accessibility on the STR$^{mut}$ haplotype not only showed significantly higher expression in the STR$^{mut}$ carriers (5.7-fold; $P = 0.0003$; two-tailed *t* test; Fig. 6a and b) but was also the most transcriptionally active region within the RTSH linkage interval. The ROI contains the stem loops of at least two miRNAs, *MIR7–2* and *MIR1179,* indicating that it corresponds to their bicistronic primary precursor (*pri-MIR*). Most *pri-MIR* transcripts terminate at a single polyAsite at the 3' end of the ROI (Extended Data Fig. 6a). Readthrough transcriptional activity that does not terminate at this polyA site was detectable at low level by RNA-seq and reverse transcription (RT) PCR. These fully spliced transcripts comprise three or four exons within the ROI that are connected to the first coding exon of the adjacent *AEN* gene. The core promoter of the *pri-MIR* lies within a CpG island and shows bidirectional activity producing the *pri-MIR* in sense, and an uncharacterized long intergenic non-coding RNA *(LINC01586)* on the opposite strand (Extended Data Fig. 6a). Here 5' RACE experiments confirmed that in both STR$^{mut}$ and NL thyroids, the identical major transcriptional start site in the core promoter was used (Supplementary Fig. 4). However, while the low-level expression of *LINC01586* and the spliced read through transcripts (as proxy for *pri-MIR* expression) in both directions are roughly similar in NL thyroid or non-RTSH diseased thyroid tissues (autoimmune thyroid disease (AITD), papillary thyroid cancer (PTC) and MNG), such concerted expression seems to be abolished in STR$^{mut}$ (Fig. 6c). Thus, STR$^{mut}$ appears to exert a directional effect on this normally bidirectional locus.

To further corroborate the relationship between STR$^{mut}$ and overexpression of the miRNA locus, we assessed whether the *pri-MIR* ROI is overexpressed from both alleles (trans) or only from the allele carrying STR$^{mut}$ (cis). Preferential allelic expression was detected for common, heterozygous SNVs within the *pri-MIR* ROI using variant calling on RNA-seq data. Segregation of the overexpressed haplotype with STR$^{mut}$ was confirmed by genotyping of the respective pedigrees (Fig. 6d). Specific overexpression of the STR$^{mut}$ containing allele was further substantiated by Sanger sequencing from the participants' genomic DNA and thyroid cDNA (Fig. 6e) and by allele-specific amplification of the spliced readthrough transcripts (Supplementary Fig. 5).

We next used small RNA-seq to define the impact of the STR genotype on the mature miRNA profile. The miRNA profiles of $STR^{mut}$ thyroids clearly differed from those with $STR^{wt}$ (either NL or MNG thyroid; Fig. 6f). In the two $STR^{mut}$ glands, MIR7–5P (the major product of *MIR7–2)* amounted to 61% and 76% of the mature miRNA pool, respectively, an estimated 10.9-fold higher level compared with NL thyroids ($FDR = 5 \times 10^{-4}$; Fig. 6g and Supplementary Fig. 6a). MIR1179 expression was also substantially induced in $STR^{mut}$ thyroid glands (2.6-fold; FDR = 0.03), but its mean expression level was comparatively low. Specific overexpression of both mature miRNAs in $STR^{mut}$ thyroids, but not skin fibroblasts (Supplementary Fig. 6b), was confirmed by real-time PCR assays including additional NL and pathological thyroid specimens (all $STR^{wt}$; Fig. 6h,i).

Other potential products of the *MIR7–2* stem loop were only expressed at negligible levels (MIR7–2-3P and MIR3529), and none of the putative unannotated miRNAs mapped to the ROI (Supplementary Table 2). Expression of the two other loci producing MIR7–5P was either undetectable (*MIR7–3*) or did not differ between $STR^{mut}$ and $STR^{wt}$ (*MIR7–1*) in thyroid (Supplementary Fig. 7). Furthermore, expression profiling in 20 normal human tissues indicated that both the spliced readthrough transcripts and MIR7–5P are preferentially expressed in thyroid tissue (Extended Data Fig. 6b). On the other hand, *MIR7–2* precursor with unprocessed stem loop was barely detectable in thyroid tissue in contrast to its detection in other tissues such as brain or small intestine (Extended Data Fig. 6b). Together, transcription of the *pri-MIR* as well as tissue-specific differences in the efficiency of its processing point to a role of MIR7–5P in NL thyroid function. Overall, these expression abnormalities support the concept that $STR^{mut}$ causes the RTSH phenotype via dysregulation of MIR7–5P and/or MIR1179 target genes in thyrocytes.

## MIR7–5P modifies proliferation pathways in $STR^{mut}$ thyroids

MIR7–5P is well established to have antiproliferative effects when overexpressed in normal or cancer cells[21] owing to its targeting of genes involved in growth factor-stimulated cell growth, such as growth factor receptors (EGFR[22,23] and IGF1R[24–26]) and the linked PI3K/AKT/MTOR signaling cascade[27,28]. For example, MIR7–5P overexpression in thyroid cell lines decreases proliferation at least in part by modulation of PI3K/AKT/MTOR activity[29]. Although TSH regulates thyrocyte differentiation, activity and growth primarily via cyclic AMP-dependent signaling, its effect on cell growth and proliferation is also mediated through the MTOR pathway[30,31], suggesting a plausible mechanism by which MIR7–5P overexpression in thyrocytes from individuals with $STR^{mut}$ may lead to RTSH.

We analyzed differentially expressed genes (DEG) to gain further insight into the pathophysiology of $STR^{mut}$ and the involvement of MIR7–5P-regulated pathways. Principal component analysis (PCA) of the RNA expression data showed that $STR^{mul}$ thyroid glands formed their own cluster separated from both NL and MNG thyroids (Fig. 7a). A total of 1,166 genes were DEG $inSTR^{mut}$ ($FDR < 0.05$ with $|\log_2(FC)| > 0.58$ versus NL; Fig. 7b). Among genes with well-defined function in thyrocyte differentiation and/or hormonogenesis, one was found among the upregulated genes (*TPO,* organification of iodide) and two were downregulated (*IYD*, recycling of iodide; *SLC2647*, iodide uptake; Supplementary Fig. 8). Although quantitation of TSHR protein was not possible due to lack of specific antibodies, it is important to note that *TSHR* mRNA was expressed normally in STR$^{mut}$ thyroids and its 3'-UTR was not substantially targeted by MIR7–5P or MIR1179 in reporter assays in vitro (Supplementary Fig. 9). We performed gene set enrichment analysis (GSEA) to test whether predicted targets of any specific miRNA (miRDB) are overrepresented among genes with reduced expression in STR$^{mut}$ thyroid glands. We found that MIR7–5P targets showed the most significant enrichment among genes with lower expression in STR$^{mut}$ thyroids providing evidence for significant transcriptome modulation by this miRNA (Fig. 7c and Supplementary Fig. 10). Several of the down-regulated DEG with MIR7–5P and/or MIR1179 target sites are integral to growth factor signaling pathways, including HPN[32], EGFR[22,23], *IGFBP5* (ref. 33), *TNS3*(ref. 34), *PIK3CB*[35] and PIK3CD[27] (Fig. 7d). We investigated the evidence for inhibition or activation of common upstream regulator genes by analyzing differential expression of downstream genes, as a way to identify specific alterations in defined functional pathways. This analysis predicted a specific increase of EGF(-like) downstream signaling in the thyroid glands of STR$^{mut}$ that was not evident in NL or goitrous thyroid glands from STR$^{wt}$ individuals (Fig. 7e and Extended Data Fig. 7).

## Discussion

We report that recurrent STR mutations in a primate- and thyroid-specific *cis*-regulatory region of *MIR7–2/MIR1179* cause increased expression of the miRNA locus and autosomal dominant RTSH. The importance of this region in the regulation of TSH has also been captured in GWAS studies, and two common single-nucleotide variants within 7 kb of this STR (rs17776563 and rs1348005; Supplementary Fig. 11) are associated with normal variability in serum TSH levels[36,17]. In fact, the identification of STR$^{mut}$ and its epigenomic effects in individuals with RTSH may shed light on the functional effects of common variants at this locus affecting TSH levels.

Most microsatellite-associated diseases are due to extreme expansion of polymorphic, intragenic STR leading to repression of gene expression, gene disruption or expression of toxic RNA and/or peptides[38]. In contrast, the $(TTTG)_4$ intergenic repeat reported herein is stable, likely due to the low number of repeat units ($n = 4$), as evidenced by the rarity of alternative alleles identified in gnomAD (Extended Data Table 1). Moreover, identification of the STR4* substitution and our epigenomic data suggest that RTSH is not caused by copy number change but instead by alterations in TF binding.

Non-coding gene regulatory variants causing Mendelian conditions often disrupt TF-binding elements within individual enhancer or insulator elements or change their location relative to a specific gene. In contrast, analysis of thyroid specimens from participants with STR$^{mut}$ demonstrates the marked activation of a pre-existing weak enhancer through alteration in the spacing of adjacent TF-binding elements directly overlapping the STR$^{mut}$. The resulting increased TF occupancy is reflected in the Fiber-seq footprint generated by the STR$^{mut}$ haplotypes on nearly all chromatin fibers. Although the full extent of the TFs mediating this effect is unknown, Fiber-seq footprinting, RNA expression data and TF motif analysis suggest that this effect likely involves F0XE1, a TF known to regulate many thyroid-specific genes[18,19]. Both STR$^{mut}$ are predicted to alter the spacing between the predicted F0XE1 binding motifs, as shown in Fig. 1d.

STR$^{wt}$ localizes in proximity of a weak enhancer within an *Alu* element. In a recent study of de novo STR generation, approximately 30% of the de novo STRs overlapped with *Alu* elements, likely in the poly(A) tail, suggesting that this is a hotspot for the creation of novel STR loci[39]. In addition, *Alu* sequences are enriched in TF-binding sites and CpG dinucleotides subject to methylation. Their preferential location near gene-rich areas supports their important role in gene regulation[40]. Studies in the human genome revealed that a gain of epigenetic enhancer marks and TF-binding motifs in *Alu* retrotransposons is positively correlated with their evolutionary age, suggesting a frequent evolutionary selection of *Alu* elements to become functional enhancers[41].

This raises the question of whether the emergence of this thyroid-specific enhancer had a role in primate evolution. The appearance of the Alu*Sx* family during primate evolution is concomitant with that of the pregnancy hormone β-human chorionic gonadotropin (hCG) gene and its serial duplication in primates[42]. Paradoxically, the latter event is associated with the selection of less potent glycoprotein hormones, including TSH[43], which may be an indication of an evolutionary conflict. hCG is known to also bind the TSHR, producing a transient increase in thyroid hormone levels during the first trimester that, when excessive, can lead to premature termination of pregnancy[44]. One can thus question whether fixation in old-world monkeys and apes of this thyroid-specific enhancer and its effect on sensitivity to TSH and hCG might also impart a protective function in the hominoid evolution.

The activated enhancer cluster created by STR$^{mut}$ selectively upregulates the bicistronic *MIR7–2/MIR1179* locus found 35 kb q-terminal of the mutation site, resulting in overexpression of both mature miRNAs in participants with STR$^{mut}$ (Fig. 6). Long-range interactions between TFs bound to the enhancer and promoter regions of *MIR7–2/MIR1179* are potentially responsible for the effect of the STR$^{mut}$ on the miRNA expression. The phenotypic spectrum caused by miRNA overexpression appears to be more complex than compensated RTSH due to *TSHR* mutations. For instance, among the 82 participants with STR$^{mut}$ in our cohort, three developed large adenomatous goiters necessitating thyroidectomy, a rate over six times higher than in the general population[13], suggesting that STR$^{mul}$ may aggravate the manifestation of proliferative thyroid disease. Based on our transcriptome analysis, we propose a working model for STR$^{mut}$ pathophysiology consisting of an antiproliferative effect of MIR7–5P overexpression in thyrocytes[29] accompanied by increased activation of EGF(-like) downstream signaling (Fig. 7f). Although *EGF* mRNA expression was unchanged in STR$^{mut}$, this would not exclude increased EGF secretion and activity. A related phenomenon has already been described in mice with thyrocyte-specific combined ablation of IGF1 and insulin receptors, both targets of MIR7–5P[24–26]. These mice are born with small thyroids that with aging develop into large goiters due to increased EGF signaling, believed to be an auto-crine feedback response to the primary signaling defect[45]. Another potential mechanism for activated EGF downstream signaling is found in the suppressed expression of the serine protease hepsin (*HPN*), a predicted MIR7–5P target (Extended Data Fig. 7c,d) that cleaves the extracellular domain of EGFR, thereby preventing receptor activation[32]. Conceivably, an activated EGF signaling pathway could promote proliferative thyroid phenotypes, for instance, in individuals harboring otherwise asymptomatic adenomatous lesions associated with somatic mutations[46].

High serum TG concentrations, despite normal *TG* expression in the thyroid glands, are another manifestation of $\mathrm{STR}^{\mathrm{mut}}$ (Fig. 3b and Supplementary Figs. 1 and 8). Possible mechanisms for this abnormality include defective polarization of thyrocytes, with misrouting of TG leading to secretion at the basolateral membrane, and leakiness of thyroid follicles. The latter could be caused by specific targeting of focal adhesion kinase by MIR7–5P[47].

While MIR7–5P overexpression was shown to substantially modulate the transcriptome in $\mathrm{STR}^{\mathrm{mut}}$ thyroids (Fig. 7c), it is not feasible to further dissect the pathophysiology using standard animal models because the miRNA cluster is incompletely conserved in rodents and the STR-controlled enhancer is primate-specific. We anticipate that recent advances in the generation of transplantable functioning human thyroid organoids[48] will enable future studies that might better define the role of this thyroid-specific enhancer of the *MIR7–2/MIR1179* locus and the pathophysiology of $\mathrm{STR}^{\mathrm{mut}}$ at the molecular level.

A practical question arises regarding whether treatment with L-T4 is necessary to reduce high serum TSH levels in individuals with $\mathrm{STR}^{\mathrm{mut}}$. Although the median FT4 concentration in individuals with $\mathrm{STR}^{\mathrm{mut}}$ is lower than that of their relatives with $\mathrm{STR}^{\mathrm{wt}}$, all but one had values within the population reference range. An argument against treatment is that it will diminish the physiologic variations of thyroid hormone levels under the control of TSH, such as diurnal, seasonal, age-dependent, and especially those during puberty and pregnancy[44,49,50]. In contrast, if the development of thyroid adenomas in individuals with $\mathrm{STR}^{\mathrm{mut}}$ is proven to be TSH-dependent, its early suppression might reduce the necessity for future surgical intervention.

Overall, $\mathrm{STR}^{\mathrm{mut}}$-mediated RTSH expands the current knowledge on the pathophysiology of thyroid hormone economy. The effect of $\mathrm{STR}^{\mathrm{mut}}$ on the chromatin landscape could not have been predicted based on conserved sequence features and epigenomic data for $\mathrm{STR}^{\mathrm{wt}}$ (refs. 51–57). Performing Fiber-seq on affected thyroid tissue was essential for the identification of an enhancer cluster activated by $\mathrm{STR}^{\mathrm{mut}}$ that causes preferential allelic overexpression of *MIR7–2/MIR1179* on the same haplotype. Over the last several years, at least five Mendelian disorders have been linked to mutations in miRNA stem loops[58–61] or deletion of a miRNA cluster[62], but $\mathrm{STR}^{\mathrm{mut}}$-linked RTSH appears to be uniquely caused by abnormal *pri-MIR* expression. We predict that non-coding mutations with similarly complex effects and mechanisms may account for a substantial number of Mendelian conditions and will be identified with increasing frequency as the use of epigenomic techniques becomes routine in gene discovery and diagnostics.

## Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-024-01717-7.

## Methods

### Human study population

Families were referred to the University of Chicago and participated in an IRB-approved protocol requiring blood samples from the proband, both parents, all siblings and children. Exceptions included death, estrangement and refusal to participate. All participants or their legal guardians provided informed written consent and did not receive compensation. From 1998 to 2022, blood samples, clinical information and thyroid tests were collected from 101 families with the phenotype of RTSH.

### WGS and analysis

Genomic DNA from two affected individuals of each of five families with chrl5-linked RTSH (Fig. 2) were subjected to WGS analysis. Library preparation and sequencing were performed either by Macrogen, Inc. (Seoul, Korea) or at the University of Washington Center for Mendelian Genomics (now the University of Washington Center for Rare Disease Research). BAM files were aligned to a human reference (hg19hs37d5) using BWA-MEM (Burrows-Wheeler Aligner; v0.7.15), and variant calling and genotyping were performed using the HaplotypeCaller tool from GATK (v4.2.0.0). GATK was used to mark sites that were of lower quality/false positives (for example, low-quality scores (Q50), allelic imbalance (ABHet 0.75), long homopolymer runs (HRun > 4) and/or low quality by depth (QD < 5)). Variants within the linkage region were annotated with Ensembl Variant Effect Predictor v83 (ref. 63) and filtered using GEMINI v0.30.1 (ref. 64) for genotype call quality (GQ ≥ 20), read depth (≥6), allele frequency in population controls (that is, maximum frequency in any continental superpopulation ingnomADv2.1 and v3.0 exomes and genomes <0.005), consistency with the mode of inheritance in each family and predicted impact on protein-coding sequence (for example, annotated as missense, nonsense, canonical splice or codingindel). For non-coding analysis, a more stringent allele frequency filter of ≤0.001 was used.

### Sanger sequencing of the STR site

To confirm the WGS findings and screen additional RTSH families for STR variant(s), one or more members from each of 71 families (including the five families investigated by WGS) were screened by Sanger sequencing. The remaining 30 families had already other defects identified using biochemical analysis or by candidate gene approach (Fig. 3a). All available members of the 12 families in which one individual had an STR variant were genotyped by Sanger sequencing. PCR primers used to amplify a fragment comprising the predicted *cis*-regulatory region surrounding the STR site are listed in Supplementary

Table 3. The PCR products were purified with ExoSAP-IT (Thermo Fisher Scientific) and sequenced at the DNA sequencing core facility of the University of Chicago.

## Thyroid function tests

TSH was measured using automated immunometric platforms, mainly Elecsys 2010 technology (Roche Molecular Biochemicals GmbH and Hitachi) and serum TG by an in-house radioimmunoassay[65]. Methodology for the measurement of FT4 varied over the period of the study and used various direct commercial immunometric assays or was estimated from the resin T4 uptake ratio and the total T4 concentrations known as the free thyroxine index (FT4I)[66]. As the units and reference ranges varied, values were rescaled to a single standard reference range (6–27 pmol $l^{-1}$) using a location-scale model[67], thereby preserving the relative position within the reference interval. Free triiodothyronine (FT3) was measured by an immunometric method at the Great Ormond Street Hospital for Children, London, UK. In vivo thyroid function was measured mainly as uptake 24 h after administration of radioactive iodide or 20 min after the administration of Tc-99m pertechnetate.

## Tissue collection

Specimens from $STR^{mut}$ thyroid glands were obtained from two participants requiring thyroidectomy because of large goiter with compression symptoms. Participant I-2 from family 30 (PtI), a male with STR 4/4*, was 59 years old and was already treated with L-T4 for 3 years when he underwent thyroid surgery for enlarged gland causing difficulty swallowing. The removed tissue weighed 131 g (eight times the normal size) and contained multiple nodules with areas of adenomatous hyperplasia and hemorrhage surrounded by a narrow rim of normal-appearing tissue. Participant II-4 from family 60 (Pt2) with STR 4/3 was 46 years old when operated for enlarged thyroid gland containing multiple hemorrhagic nodules. A total of 156 g (9.7-fold the normal size) was removed. Biopsies of healthy thyroid tissue (NL; $STR^{wt}$) were collected from seven participants requiring surgery for parathyroid adenomas. Additional samples of diseased thyroid tissue (all with $STR^{wt}$) were obtained from one participant with non-toxic MNG, four participants with goiter due to AITD and two participants with PTC. All thyroid specimens were immediately snap-frozen and stored at −80 °C before processing. Human skin was obtained by punch biopsy from three participants with $STR^{mut}$ and five control individuals with $STR^{wt}$. Dermal fibroblast cultures were maintained in DMEM supplemented with 10% FBS[68]. All studies were approved by the Institutional Review Board of the University of Chicago.

## Tissue CpG methylation analysis

Genomic DNA isolated from thyroid glands, cultured skin fibroblasts, or leukocytes was subjected to bisulfite treatment (EZ DNA Methylation-Gold Kit, Zymo Research). Subsequently, PCR amplification was performed using primers designed to specifically amplify the $STR^{wt}$ and $STR^{mut}$ haplotypes and introduce methylation-sensitive restriction endonuclease sites for CpG methylation analysis (sequences of primers available upon request). The digested PCR products were resolved on agarose gels and relative methylation was determined by image quantification (ImageJ). Each experiment included positive and negative methylation controls (Zymo Research).

## Fiber-seq sample processing and sequencing

Thyroid tissue samples from a healthy control individual (NL;STR$^{wt}$), an individual with MNG (STR$^{wt}$) and two participants with RTSH (one with STR3 and another with STR4* variant) were processed. Specifically, 150–200 mg of frozen tissue was minced using a razor blade and then placed in 1 ml of cold homogenization buffer (250 mM sucrose, 10 mM Tris-CL, pH 8.0, 25 mM KCI, 5 mM $MgCl_2$, 0.1 mM DTT, 0.2 U RNasin Plus (Pro-mega, N2615), 0.1% Triton X-100 and 1× protease inhibitor (Promega, G6521)). Tissue was then homogenized using a Dounce homogenizer with ten strokes using pestle A, followed by ten strokes using pestle B. An additional 1 ml of cold homogenization buffer was added to the dounced sample, which was then filtered through a 70-micron filter. The samples were then centrifuged at 350*g* for 10 min and the supernatant was removed. The pellet was then resuspended in 360 µl of buffer A (15 mM Tris-CI, pH 8.0, 15 mM NaCI, 60 mM KCI, 1 mM EDTA, 0.5 mM EGTA, 0.5 mM spermidine) and divided equally into six PCR tubes. To each tube, we added 1.5 µl of 32 mM S-adenosylmethionine, as well as 100 units of Hia5 enzyme[14] and incubated the sample for 10 min at 25 °C. The reactions were stopped by adding 3 µl of 20% sodium dodecyl sulfate followed by hand mixing of the tubes. The separate reactions were then combined and subjected to DNA extraction using the Wizard HMW DNA Extraction Kit (Promega, A2920). DNA was sheared using a Megaruptor device, subjected to PacBio SMRTbell library preparation, size selected and sequenced using a PacBio Revio sequencer[16].

## Fiber-seq data processing and peak calling

PacBio Revio BAM files containing kinetic information were initially processed using jasmine (v2.0.0) to identify methylated CpGs and then by fibertools (v0.2.6)[15,69] to identify per-molecule m6A-modified bases, nucleosome footprints and methyltransferase-accessible patches (MSPs) that represent methylated stretches between nucleosome footprints. Reads were then haplotype-phased using a custom pipeline (*k*-mer-variant-phasing v0.0.l)[70] that first identifies SNVs using Deep-Variant (vl.5.0)[71], and then runs a variant-based phaser (HiPhase v0.10) to bin reads into phase blocks[72]. These phase blocks are then assigned to either the maternal or paternal haplotype using parental short-read genome sequencing combined with the trio *k*-mer-based phaser HiCanu[73]. Haplotype-phased single-molecule m6A-modified reads are then processed using the Fiber-seq Inferred Regulatory Element (FIRE) pipeline (FIRE v0.0.0-alpha.1 (ref. 74); Vollger et al., *in preparation)*. Briefly, the FIRE pipeline using a semisupervised machine-learning approach[75,76] trained using a mixed positive dataset (that is, features of MSPs overlapping known DNaseI hypersensitive sites and CTCF ChIP-seq peaks) and negative dataset (that is, features of MSPs within the mappable genome that do not overlap known DNaseI hypersensitive sites or CTCF ChIP-seq peaks). This produces a score for each MSP with lower scores corresponding to MSPs classified as matching accessible chromatin elements and higher scores representing MSPs classified as internucleosomal linker regions. Aggregate Fiber-seq measurements of chromatin accessibility are generated by summing the -$\log_{10}$ of these scores for MSPs overlapping each position along the genome. Accessible chromatin peaks are then identified at a genome-wide Bonferroni-corrected significant threshold of 0.01. Peaks with significant differences in accessibility between haplotypes are identified by counting the number of fibers from each haplotype that have a FIRE MSP with a score of <0.05 versus >0.05 and calculating a Fisher exact test using these values. Single-molecule CpG methylation along each Fiber-seq read was identified using primrose (v1.3.0) and aggregated across reads using pb-CpG-tools (v2.3.1).

## Fiber-seq enhancer codependency

We performed enhancer cluster codependency analyses by calculating codependency scores for every pair of FIRE peaks within the enhancer cluster. We define codependency scores as the observed rate of co-accessibility (that is, accessible at both peaks along an individual chromatin fiber) minus the expected rate of co-accessibility given independence between the two peaks. Specifically, we identified overlapping chromatin fibers and accessible patches (FDR ≤ 0.05) using bedtools intersect (v2.30.0) and calculated the proportion of fibers that are accessible at each peak. We performed all subsequent codependency analyses in Python (v3.9.12) using the standard library. For each pair of FIRE peaks, we calculated the expected co-accessibility as the product of their accessible proportions, while the observed co-accessibility was calculated as the proportion of fibers spanning both peaks that are accessible at each.

We quantified the essentiality of each peak in the enhancer cluster by constructing codependency graphs, with nodes representing enhancer peaks and edge weights representing codependency scores. We constructed graphs that omitted individual peaks and limited each analysis to fibers inaccessible at that respective peak. We also constructed a baseline codependency graph containing all peaks. We quantified the total codependency of each graph by summing all edge weights and normalizing by the number of edges. Finally, we calculated essentiality as the ratio of total codependency of the baseline graph over peak-excluded graphs, with a higher ratio indicating a larger reduction in codependency following the loss of accessibility at that peak.

## High-throughput RNA sequencing

RNA samples were prepared from frozen thyroid tissue samples using Qiagen miRNA easy/ Rneasy MinElute Cleanup protocol to produce separate small RNA (<200 nt) and long RNA (>200 nt) fractions. For the long RNA fractions, ribosomal RNA was depleted by subtractive hybridization (NEB Next rRNA Depletion Kit, NEB) and libraries constructed with TruSeq Stranded Total RNA Library Prep (Illumina). Sequencing was performed on an Illumina HiSeq 4000 system (paired-end, 150 cycles). The libraries for the small RNA fractions were prepared using NEB Next Small RNA Library Prep Set (NEB) and sequenced (single end, 50 cycles) on an Illumina HiSeq 4000 system. The quality of the raw reads was verified, for each sample, using FastQC (v0.11.3; http://www.bioinformatics.babraham.ac.uk/projects/fastqc).

## Analysis of RNA-seqdata

For the paired-end RNA-seq data, the Tuxedo Suite software package was used for alignment, differential expression analysis and postanalysis diagnostics (http://gitluib.com/umich-brcf-bioinf/Watermelon/releases/tag/release-v0.3.4 (refs. 77–79)). Reads were aligned to UCSC hgl9 using TopHat (v2.0.13) and Bowtie2 (v2.2.1), using the default parameter settings for alignment, except for the preset Bowtie2 option ‘b2-very-sensitive’ for the most sensitive and accurate alignments. The following two different techniques were used for quantitation, normalization and differential expression analysis: (1) Cufflinks/ Cuffdiff (v2.1.1), with parameter settings: '--multi-read-correct' to adjust expression calculations for reads that map in more than one locus, as well as '--compatible-hits-norm' and '--upper-quartile-norm' for normalization of expression values. (2) HTSeq (v0.6.1), to count non-ambiguously mapped reads only. Data were prefiltered to remove genes with 0 counts in all samples. Normalization and differential expression were then performed with DESeq2 (vl.14.1), using a negative binomial generalized linear model. For the small RNA-seq data, we used the CAP-miRSeq pipeline (http://bioinformaticstools.mayo.cdu/research/cap-mirseq/ (ref. 80)), human reference genome version UCSC hgl9 and MirBase (v21). For both long-RNA and microRNA, differentially expressed transcripts were identified based on FDR ≤ 0.05 and $|\log_2(\mathrm{FC})| > 0.58$. Coverage charts for the mapped reads and Sashimi plots were created in Integrative Genomics Viewer[81], and regional read counts were calculated with SAMtools[82].

## Analysis of preferential allelic expression

We genotyped common SNVs (MAF > 0.2) in a 5 kb region surrounding the *MIR7–2* stem loop from gDNA samples corresponding to the thyroids analyzed by RNA-seq, to identify a set of SNVs that were heterozygous in both participants and at least two of the healthy controls. Variant calling on RNA-seq data was performed using a Sentieon optimized analysis pipeline (Basepair, Inc.) that follows the GATK best practices for short variant calling on RNA-seq data. Steps included read mapping to the hg38 reference genome using STAR in the two-pass mode, removal of duplicate reads, splitting of reads at junctions, base quality score recalibration and variant calling through the Haplotyper algorithm with the minimum phred-scaled confidence threshold (call_conf, emit_conf) set to 20. Only SNVs with filtered depth (DP) > 20 were retained for analysis ($\mathrm{STR}^{\mathrm{mut}}$: DP = 213 ± 22 SEM; normal: DP = 48 ± 14 SEM). In another test for preferential allelic expression, we performed allele-specific real-time 'RT-PCR of the spliced readthrough transcript using allele-specific reverse primers for rs3743476 (MAF = 0.33) in the first coding exon of *AEN*.

## Real-time RT-qPCR

Deoxyribonuclease-treated total RNA preparations from thyroid tissue or cultured skin fibroblasts were reverse transcribed using random hexamer priming[83]. Real-time PCR was then performed with SYBR Green dye (Life Technologies) and gene expression was calculated using the $2^{-\Delta\Delta\mathrm{Ct}}$ method as described previously[83]. Primer sequences are provided in Supplementary Table 3. The reference gene set (*ACTB*, *GAPDH, PPIA* and *IPO8)*[84] was confirmed to be stably expressed in the TMM-normalized RNA-seq data from $\mathrm{STR}^{\mathrm{mut}}$ and $\mathrm{STR}^{\mathrm{wt}}$ thyroid glands. *UBE2D3* (ref. 85) was used as a reference for normalization across different types of tissues. The expression of mature miRNAs was measured using predesigned TaqMan real-time PCR assays (Applied Biosystems and Thermo Fisher Scientific)[86] specific for human MIR7–5P, MIR1179 and the reference snoRNA RNU44 (also known as SNORD44).

## Gene set enrichment and upstream regulator prediction

To test whether specific miRNAs substantially modulate the thyroid transcriptome in STR$^{mut}$, a gene list preranked by $\log_2$(FC) (STR$^{mut}$ versus normal) was used in GSEA (v4.3.2; http://www.gsea-msigdb.org/gsea/index.jsp (ref. 87)) against 2,141 miRNA target gene sets (size limit: 15–500 genes per set) from the Molecular Signature Database (c3.mir. v2023.1.Hs.symbols.gmt)[88]. Upstream regulators of the DE genes were predicted using the proprietary iPathwayGuide tool (http://advaitabio.com/bioinformatics/ipathwayguide/ (ref.S9)).

## 5’RACE

The GeneRacer (Invitrogen) protocol for RNA-ligase-mediated, oligo-capping rapid amplification of cDNA ends (RLM-RACE) was used for mapping of transcription stare sites[90,91]. This method results in the selective ligation of an RNA oligonucleotide to the 5’ ends of decapped mRNA using T4 RNA ligase ensuring that only full-length transcripts are amplified. The gene-specific primer (Supplementary Table 3) was designed to map to an exon of the readthrough transcript expressed from the *pri-MIR* promoter region.

## *TSHR* 3’-UTR luciferase reporter assay

To create the *TSHR* 3’-UTR luciferase reporter construct, the *TSHR* 3’-UTR (chrl4: 81,144,603–81,146,145; hg38) was inserted into pGL3-promoter (between the firefly luciferase coding sequence and an SV40-poly(A) cassette). Clones with reverse orientation of the 3’-UTR were isolated as additional controls. The *pri-MIR7–2* and *pri-MIR1179* expression plasmids were obtained by cloning ~500 bp fragments comprising the respective stem loop sequences into pcDNA3. All constructs were confirmed by sequencing. Reporter constructs (250 ng cm$^{-2}$ of cell monolayer) were cotransfected in HEK293 (ATCC, CRL-1573) or COS1 (ATCC, CRL-1650) cells with empty pcDNA3 or expression plasmids for either *pri-MIR7–2* or *pri-MIR1179* (each at 106 ng cm$^{-2}$) using FuGENE 6 reagent (Promega). All transfection mixes included a spike-in of *Renilla* luciferase plasmid (pRL-TK; at 8 ng cm$^{-2}$) for normalization. Firefly and *Renilla* luciferase activities were measured 24 h post-transfection (Dual Luciferase Assay system, Promega).

## Statistics and reproducibility

No statistical method was used to predetermine sample size and no randomization was used as this study involved genetic testing of individuals with a rare monogenic disease. No data were excluded from any of the experiments described. Data collection and analysis were not performed blind to the phenotype or genotype of the study participants. Genomic analyses (WGS, RNA-seqand Fiber-seq) of samples were performed once, while PCR-based expression data represent assay results from at least two technical replicates for the shown set of samples. All attempts at replication were successful. As indicated, we evaluated group differences for statistical significance using two-tailed Student’s *t* test or, if at least one group did not pass the D’Agostino-Pearson normality test, by the two-tailed Mann-Whitney test. Data including very small group sizes ($n \leq 5$) were assumed to follow a Gaussian distribution with equal variance, but this was not formally tested. Data were analyzed with GraphPad Prism (vl0.0.0). *P* values were adjusted for multiple testing with the Benjamini-

Hochberg procedure to control the FDR. The FDR for the association of STR$^{mut}$ with RTSH was estimated using a binomial cumulative distribution function[92]. Results with a value of $P < 0.05$ were considered statistically significant.

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

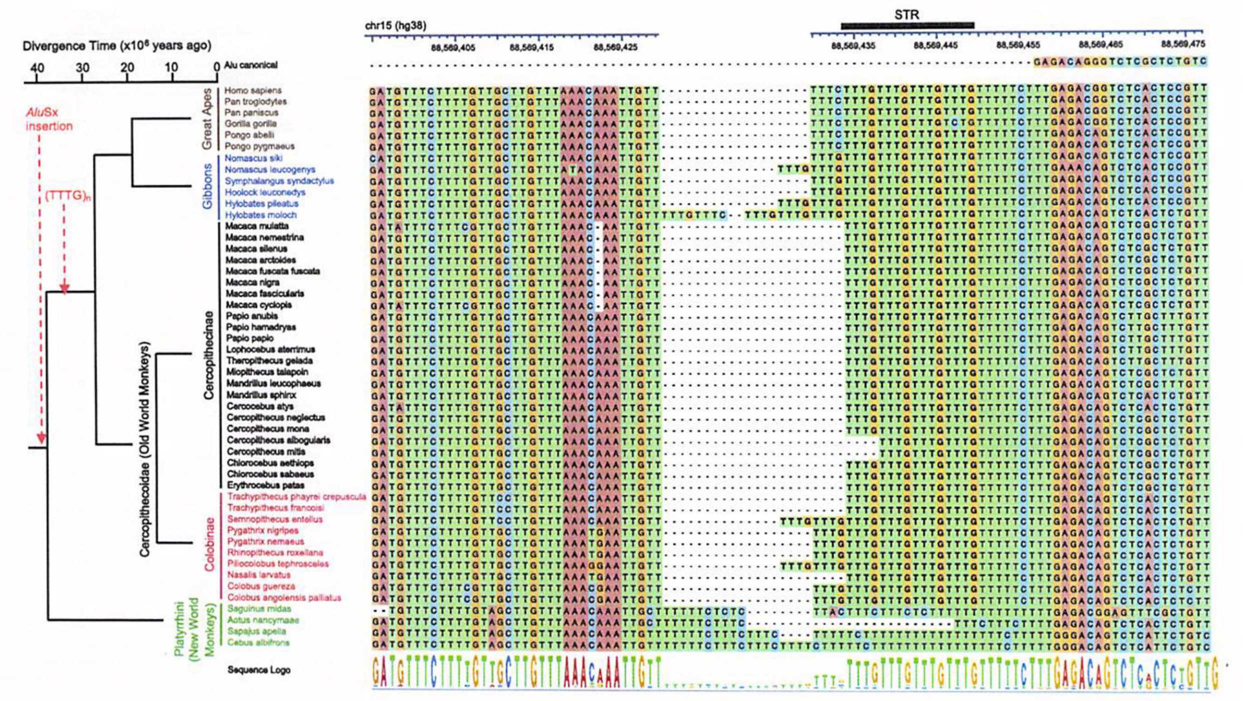
## Data availability

All sequence data from the RNA-seq studies are available from the NCBI Sequence Read Archive (BioProject PRJNA972324) at http://www.ncbi.nlm.nih.gov/sra/PRJNA972324. In situ Hi-C contact matrix of the human Embryonic Stem Cell line H1 (ref. 93) was obtained from the 4D Nucleome Data Portal at http://data.4dnucleome.org/. The thyroid-specific ChIP-chip data for histone modifications were from the International Human Epigenome Consortium (IHEC) data portal (http://epigenomesportal.ca/ihec/grid.html), as well as ENCODE (http://www.encodeproject.org/; entries ENCSR033KMZ, ENCSR749MUH, ENCSR203KCB, ENCSR906YES, ENCSR975NOU and ENCSR432GAO). snATAC-Seq data[57] were downloaded from the Human Cell Atlas data portal at http://data.humancellatlas.org/explore/projects/c31fa434-c9ed-4263-a9b6-d9ffb9d44005. RNA-seq and microRNA-seq data was downloaded from ENCODE (ENCSR687HJY and ENCSR566RDG). Gene sets for GSEA were obtained from the Molecular Signature Database at http://data.broadinstitute.org/gseamsigdb/msigdb/release/2022.1.Hs/. miRNA target gene sets were obtained from miRDB (v6.0; http://mirdb.org/download.html)[88] and MicroT-CDS (http://dianalab.e-ce.uth.gr/tools)[94]. Restrictions apply to the availability of Fiber-seq, WGS and patient-specific data generated or analyzed during this study to preserve participants' confidentiality. The corresponding author will on request detail the restrictions and any conditions under which access to some data may be provided. Source data are provided with this paper.

## Code availability

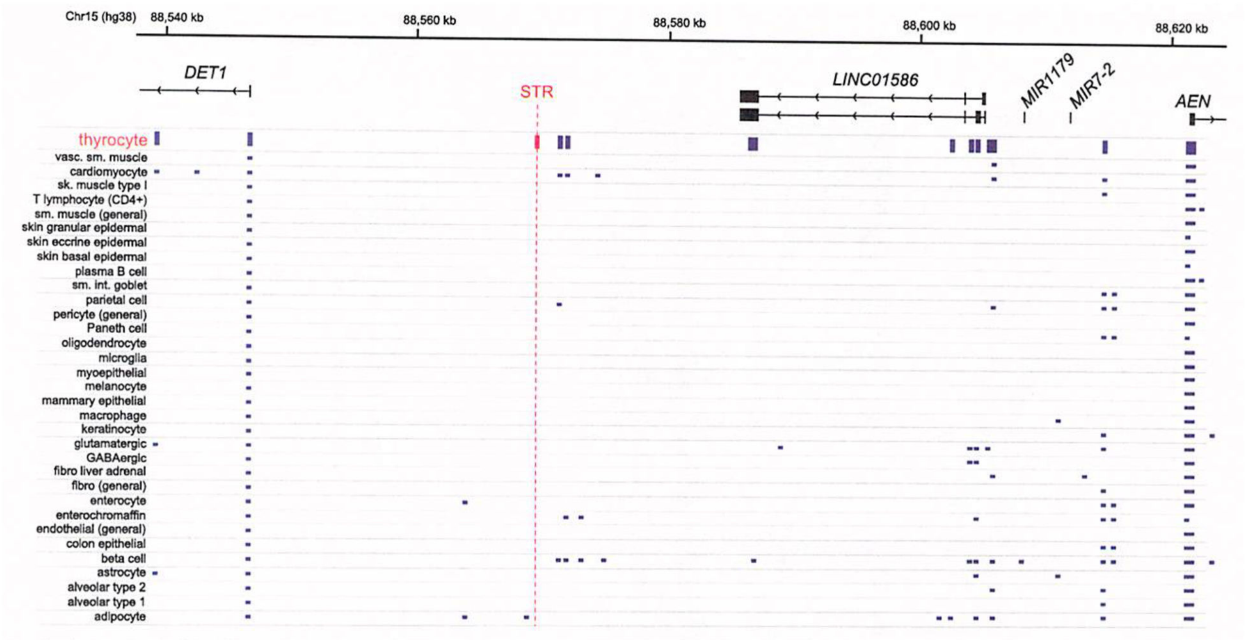
All software used in the study are publicly available as described in the Methods and Reporting Summary. All custom codes used to analyze the Fiber-seq data and generate plots for this study have been deposited in an open repository[69,70,74,95].

# Extended Data



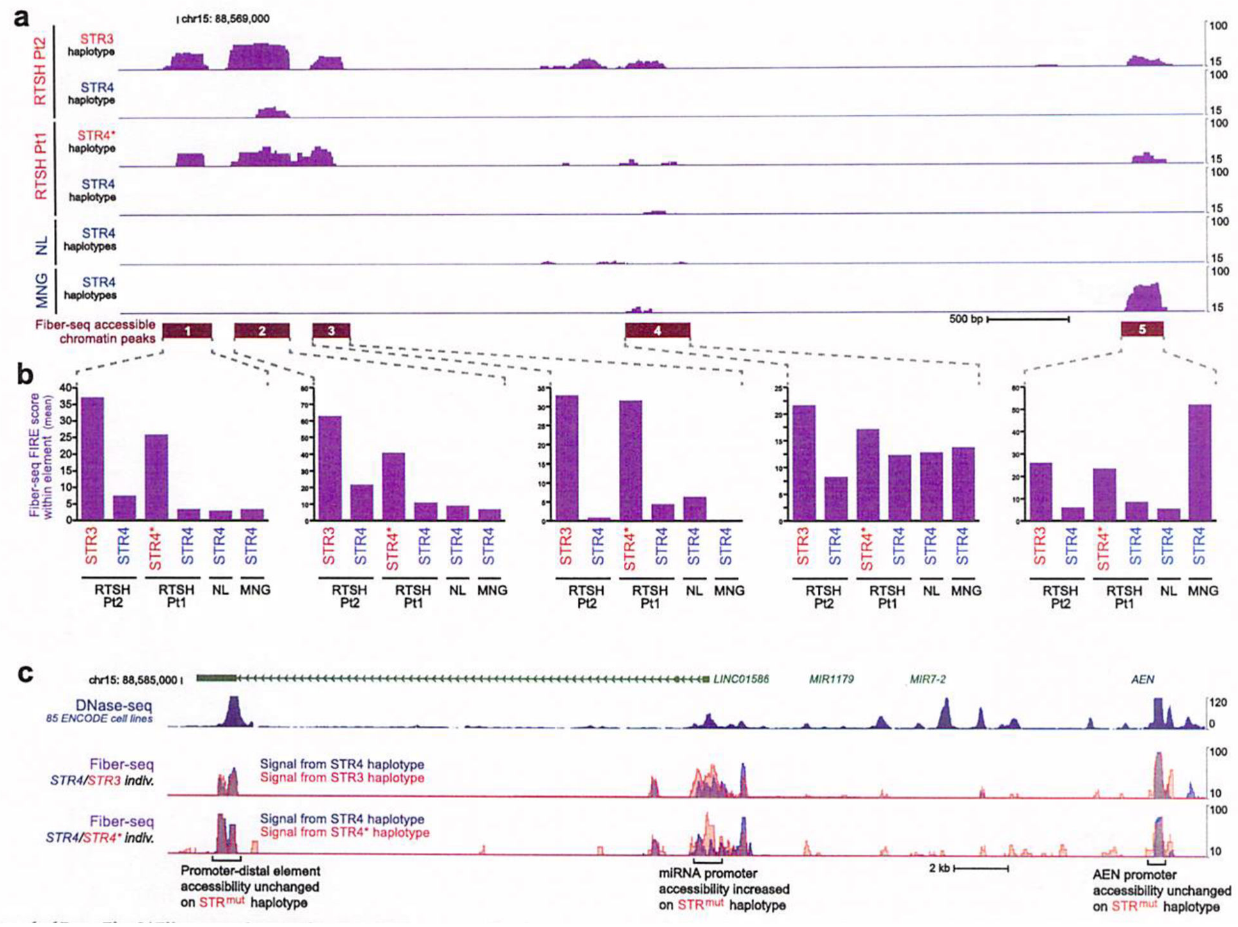
**Extended Data Fig. 1 |. Alignment of the human STR $(TTTG)_4$ with homolog sequences from other primates.**

We searched 86 nonhuman primate genomes available in the NCBI databank (refseq_genomes; accessed 09/2023) by BLAST search with a 278 bp human sequence centered around the STR. Of 49 Old World Primates, 47 had a single match over the full sequence that was confirmed to be syntenic (that is, upstream of the *MIR7–2* locus). All of these had a TTTG repeat sequence within the T-rich sequence (corresponding to the poly(A) tail of the *Alu*Sx1 retrotransposon). For the remaining two Old World Primate genomes *(Rhinopithecus strykeri* and *Rhinopithecus bieti),* we were not able to ascertain a clear homolog, although the homolog was present in another member of the same genus (*R. roxellana)*. For the 37 other nonhuman primate genomes (including 13 New World Primates, 2 Tarsiiformes, 1 Chyromyiformes, 16 Lemuriformes, 5 Lorisiformes), only 4 of the New World Primate genomes *(Cebus albifrons, Sapajus apella, Sanguinus midas, Aotus nancymaae)* had a detectable syntenic homolog of the STR encompassing region, but their T-rich sequence was notably devoid of TTTG repeats. These data place the insertion of the *Alu*Sx1 retrotransposon into this locus before the split between New and Old World Primates, that is about 40 million years ago. This finding is consistent with data showing that the *AluS* subfamily arose from *Aluj* after the split between Strepsirrhini (Lemuriformes and Lorisiformes) from the common ancestors of Old and New World Primates[96,97]. Since essentially all extant members of the Old World Primate lineage seem to carry a TTTG repeat sequence in this interval, we postulate that it first evolved in the common ancestor of Old World Primates. Note that for genera with multiple representatives with virtually identical sequences, only a subset of species is shown.



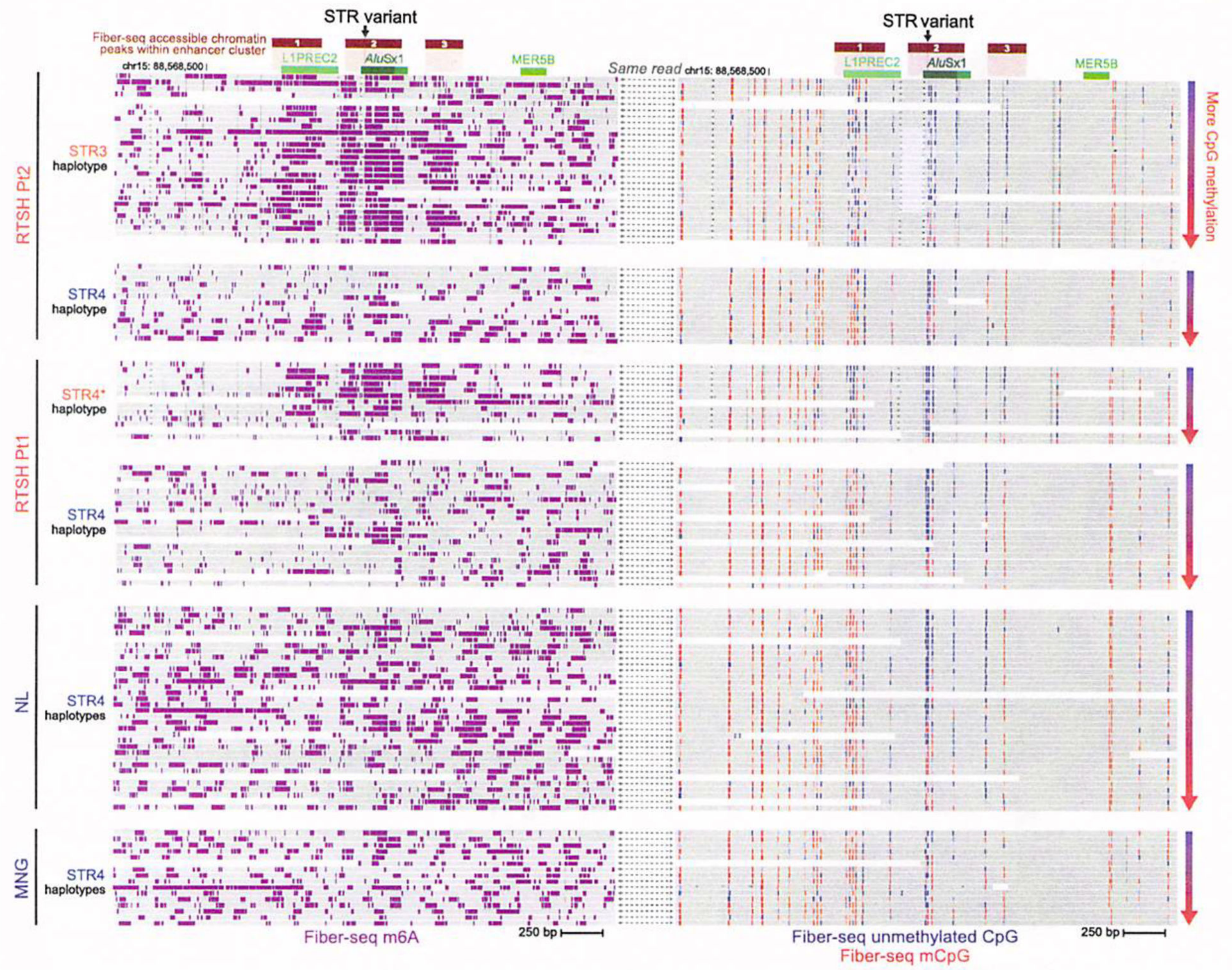
**Extended Data Fig. 2 |. Thyroid-specificity of the STR-associated *cis*-regulatory element.**

Shown are snATAC-Seq peak data[57] for different cell types. The STR maps to a predicted *cis*-regulatory element specifically active in thyroid follicular cells.



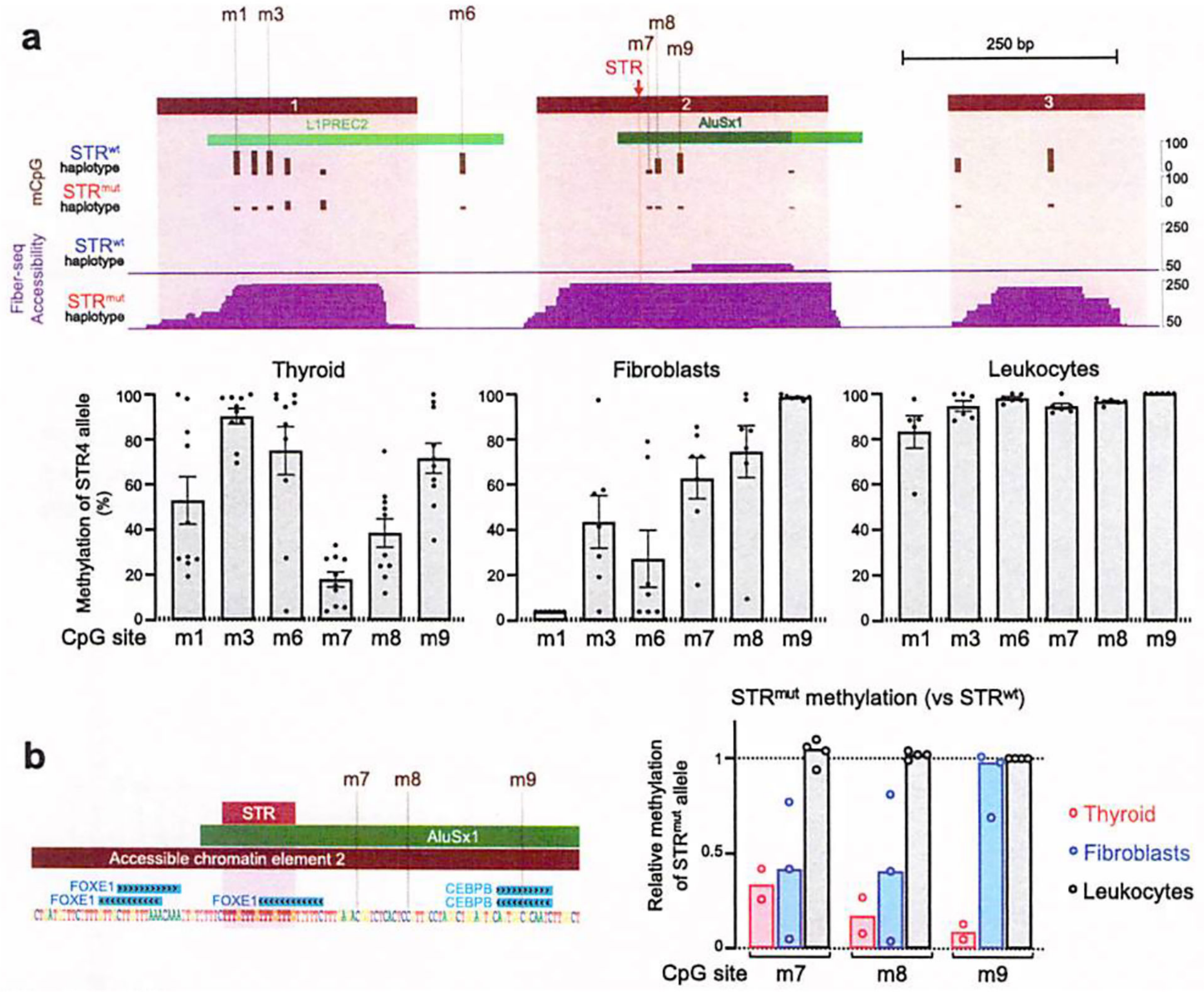
**Extended Data Fig. 3 |. Fiber-seq chromatin accessibility at the STR site and miRNA locus.**

**a,** Fiber-seq chromatin accessibility tracks showing the FIRE scores at the enhancer cluster surrounding the STR site in thyroid tissue from two individuals with RTSH, thyroid tissue from a healthy control individual, and MNG thyroid tissue. For the individuals with RTSH, FIRE signal is separated by haplotype. **b,** Bar plots showing the mean FIRE scores within the elements highlighted above across the samples shown in **a.** Of note, multi-nodular goiter (MNG) tissue showed selective activation of element 5 within this enhancer cluster, with elements 1–3 remaining silenced. **c,** DNase-seq signal at the miRNA locus from ENCODE, as well as haplotype-specific Fiber-seq chromatin accessibility in this region. Note that the miRNA promoter region selectively demonstrates haplotype-specific chromatin accessibility.



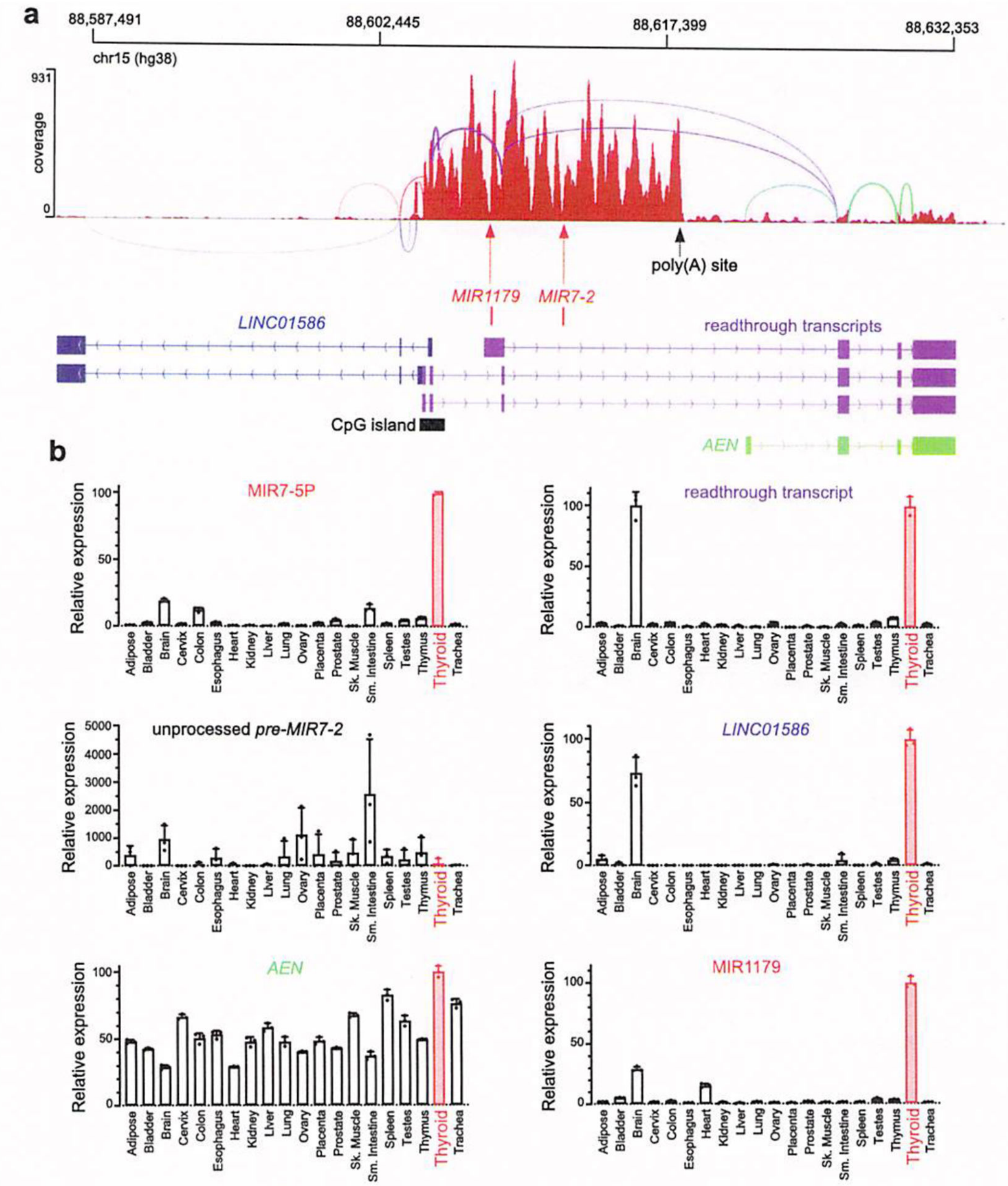
**Extended Data Fig. 4 |. Relationship between CpG methylation and chromatin accessibility at the STR enhancer cluster.**

For each Fiber-seq sequencing read at this site, shown is the per-molecule m6A-marked chromatin accessibility stencils (left), as well as the per-molecule methylation status of CpG dinucleotides (right) for the same reads. CpG methylation status is indicated using a color gradient from blue to red, with blue indicating unmethylated CpGs, and red indicating methylated CpGs, and the shading according to the precision of the methylation calls according to Primrose. Reads are separated out by sample, as well as by haplotype within the RTSH samples. Reads are then ranked based on the CpG methylation status surrounding the STR site, with hypo-CpG methylated reads being on top. This reveals that individual chromatin Fibers harboring chromatin accessibility at this enhancer cluster similarly had hypo-CpG methylation of the surrounding CpG dinucleotides. However, hypo-CpG methylation of the surrounding CpGs was similarly observed on fibers lacking chromatin accessibility at elements 1–3, indicating that hypo-CpG methylation is necessary, but not sufficient for actuating this codependent enhancer cluster.



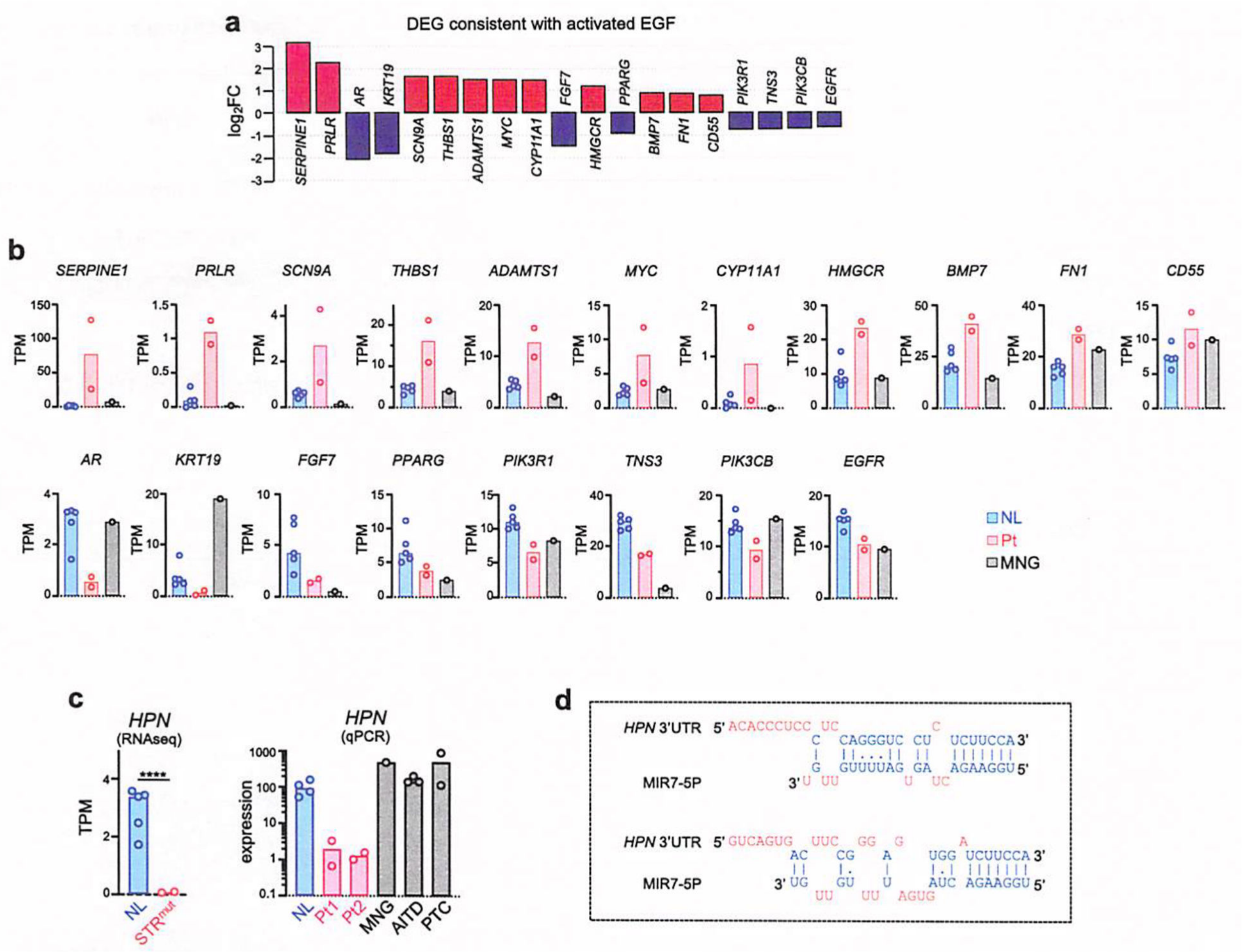
**Extended Data Fig. 5 |. Tissue-specific DNA hypomethylation of the STR$^{mut}$ allele.**

**a,** Upper panel depicts selected CpG sites hypomethylated on the STR$^{mut}$ haplotype in thyroid (compare Fig. 1a). The tissue-specific methylation profiles of the STR$^{wt}$ allele were established by allele-specific amplification from bisulfite treated DNA from either thyroid tissue (from n = 10 individuals per CpG site), skin fibroblasts (n = 7) or leukocytes (n = 6). and the methylation status (%methylated; mean±SEM) of six CpG sites (m1, m3, m6, m7, m8, m9) quantified following digestion with methylation-sensitive restriction endonucleases. Three CpG sites (m7, m8, m9) close to the STR$^{wt}$ sequence were hypomethylated in thyroid compared to fibroblasts and leukocytes suggesting their relevance for thyroid-specific expression under normal conditions. **b,** Tissue-specific relative methylation of STR$^{mut}$ (vs STR$^{wt}$) in thyroid tissue (from n = 2 participants with RTSH), cultured fibroblasts (n=3), and leukocytes (n = 4). On the chromosome harboring STR$^{mut}$, m7 and m8 appear to be hypomethylated in fibroblasts albeit to lesser degree than in thyroid. In contrast, for the m9 site, hypomethylation of STR$^{mut}$ chromosomes appears to be restricted to thyroid. The latter site locates to a binding motif of C/EBPB and is hypermethylated in fibroblasts (both, STR$^{wt}$ and STR$^{mut}$). These results could indicate that in fibroblasts, binding of a forkhead domain TF (other than FOXE1) at the STR region produces a limited increase in accessibility that does not extend to m9. In thyroid, FOXE1 produces more extensive accessibility followed by hypomethylation and binding of an additional TF, thatls.C/EBPB.



**Extended Data Fig. 6 |. Transcriptional activity of the *MIR7–2/MIR1179* locus indicating high thyroid specificity of mature miRNA expression.**

**a,** Top panel: Sashimi plot visualizing coverage and splice junctions from aligned RNA-seq data from Pt1. The acute absence of reads overlapping with the localization of the miRNA stem loop sequences suggests efficient processing of the *pri-MIR* by the Microprocessor complex. Lower panel: annotation track showing the relative positions of relevant features including the *MIR7–2 and MIR1179* stem loops, the major poly A site where most *pri-MIR* transcripts terminate, and spliced readthrough transcripts that connect to the first coding exon of *AEN. LINC0158* is expressed on the opposite strand and overlaps with *pri-MIR* transcripts within the core of the bidirectional promoter (containing a CpG island). **b,** Tissue expression profiling of products from the miRNA locus. For each tissue, total RNA pooled from at least 3 individual donors was utilized (Ambion). Data (mean±SD; 3 replicate amplifications) were obtained by real-time reverse transcription PCR assays, normalized for either RNU44 (MIR7–5P and MIR1179)or *UBE2D3* (spliced readthrough transcript,unprocessed *MIR7–2* stem loop, *LINC01586,AEN)*[85],.and expressed relative to the expression detected in thyroid tissue (set to 100%).



**Extended Data Fig. 7 |. Data supporting activated EGF(-like) signaling activity.**

**a,** Subset of DEGs (n = 2 STR$^{mut}$ vs n = 5 NLSTR$^{wt}$ thyroids) that were directionally consistent with EGF signaling activity (from the analysis shown in Fig. 7e). **b,** Expression values of DEGs shown in **a.** Gene expression profile of an STR$^{wt}$ (nontoxic) MNC is shown for comparison. **c,** Downregulation of *HPN* encoding an EGFR-inactivating protease. Specific suppression of *HPN* expression in STR$^{mut}$ thyroid glands by RNA-seq (n = 2 thyroids with STR$^{mut}$ and 5 NL with STR$^{wt}$) and RT-qPCR. For RT-PCR, samples Included n = 4 NL thyroids (2 distinct from RNAseq samples), n=3 with AITD. n = 2 with PTC, and one with MNG (all STR$^{wt}$). For both Pt1 and Pt2 with STR$^{mut}$, two separate specimens from their excised thyroid glands were analyzed by qPCR.****, FDR = 2 × $10^{-21}$ (DESeq2). **d,**

Alignment of MIR7–5P to predicted 3′-UTR target sites[94] of *HPN*. Watson-Crick pairings are shown as vertical dashes and G:U wobble pairings by dots. TPM, transcripts per million.

### Extended Data Table 1|

Genetic variants in gnomAD that alter the $(TTTG)_4$ repeat or the immediately flanking sequence

| Pos (chr15) | rsID | Ref | Alt | Allele Count | Allele Number | relative to $(TTTG)_4$ start/end | Note |
|---|---|---|---|---|---|---|---|
| 88569429 | rs1456817951 | TTTC | T | 172 | 152082 | pos −5 | deletion of TTTC preceeding $(TTTG)_4$ |
| 88569433 | rs965243764 | C | CTTTG | 5 | 151588 | within STR | insTTTG (“STR5”) |
| 88569433 | rs965243764 | CTTTG | C | 1 | 151588 | within STR | **delTTTG (STR3)** |
| 88569433 | rs775472969 | C | G | 3 | 151588 | pos −1 | |
| 88569436 | rs1325867852 | TG | T | 1 | 152174 | within STR | deletion of G of 1 st TTTG |
| 88569450 | rs560687645 | T | C | 1 | 152170 | pos +1 | |
| 88569452 | rs958033426 | T | A | 1 | 152196 | pos +3 | |
| 88569453 | rs989905631 | T | A | 5 | 152.194 | pos +4 | |

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgements

We thank the families for their participation and invaluable contributions to this research. We thank the following physicians for the referral of families with RTSH that turned out to have mutations in the STR on chromosome 15: I. Rosenthal, The University of Illinois in Chicago, IL, USA (deceased); R. Ruvalcaba, University of Washington, Seattle, WA, USA; M. Christine Vantyghem, Centre Hospitalier Universitaire de Lille, France; M. Abramowicz, Université Libre de Bruxelles, Belgium; G. Van Vliet, Université de Montréal, Quebec, Canada; H. Abdul-Latif, Johns Hopkins University School of Medicine, Baltimore MD, USA; D. L. Metzger, University of British Columbia, Vancouver, BC, Canada and W. Patrick Zeller, Academic Endocrine, Metabolism and Nutrition Group, Wheaton, IL, USA. We also thank S. Cheroux for help with the analysis of WGS data generated by Macrogen, S. Neph for help with the analysis of the Fiber-seq data and T. Karrison for assistance with the statistical analyses. We also thank the Great Ape Information Network, Primate Research Institute and the Wildlife Research Center of Kyoto University, JMC Kyoto City Zoo, and Ueno Zoological Gardens for providing *Gorilla* samples. This work was supported by the Cooperative Research Program of Primate Research Institute, Kyoto University (2020-C-12). This work was supported in part by grants from the National Institutes of Health (DK15070 (to S.R.), DK134486 (to H.G.), 1DP50D029630 (to A.B.S.), U01 HG011744 and UM1 HG006493 (to M.B.)), the Belgian Kids Fund (to G.S.) and not for profit Association Recherche Biomedicale et Diagnostic (to G.V. and S.C.). A.B.S. holds a Career Award for Medical Scientists from the Burroughs Wellcome Fund and is a Pew Biomedical Scholar.

## References

1. Peters C, van Trotsenburg ASP & Schoenmakers N Diagnosis of endocrine disease: congenital hypothyroidism: update and perspectives. Eur. J. Endocrinol 179, R297–R317 (2018). [PubMed: 30324792]
2. Boersma B, Otten BJ, Stoelinga GB & Wit JM Catch-up growth after prolonged hypothyroidism. Eur. J. Pediatr 155, 362–367 (1996). [PubMed: 8741031]

3. Tenenbaum-Rakover Y et al. Long-term outcome of loss-of-function mutations in thyrotropin receptor gene. Thyroid 25, 292–299 (2015). [PubMed: 25557138]

4. Persani L et al. Genetics and phenomics of hypothyroidism due to TSH resistance. Mol. Cell. Endocrinol 322, 72–82 (2010). [PubMed: 20083154]

5. Stanbury JB, Rocmans P, Buhler UK & Ochi Y Congenital hypothyroidism with impaired thyroid response to thyrotropin. N. Engl. J. Med 279, 1132–1136 (1968). [PubMed: 4176719]

6. Parmentier M et al. Molecular cloning of the thyrotropin receptor. Science 246, 1620–1622 (1989). [PubMed: 2556796]

7. Sunthornthepvarakul T, Gottschalk ME, Hayashi Y & Refetoff S Brief report: resistance to thyrotropin caused by mutations in the thyrotropin-receptor gene. N. Engl. J. Med 332, 155–160 (1995). [PubMed: 7528344]

8. Xie J et al. Resistance to thyrotropin (TSH) in three families is not associated with mutations in the TSH receptor or TSH. J. Clin. Endocrinol. Metab 82, 3933–3940 (1997). [PubMed: 9398691]

9. Grasberger H et al. Autosomal dominant resistance to thyrotropin as a distinct entity in five multigenerational kindreds: clinical characterization and exclusion of candidate loci. J. Clin. Endocrinol. Metab 90, 4025–4034 (2005). [PubMed: 15870119]

10. Grasberger H et al. Identification of a locus for nongoitrous congenital hypothyroidism on chromosome 15q25.3–26.1. Hum. Genet 118, 348–355 (2005). [PubMed: 16189712]

11. Grasberger H & Refetoff S Resistance to thyrotropin. Best Pract. Res. Clin. Endocrinol. Metab 31, 183–194 (2017). [PubMed: 28648507]

12. Aliesky H, Courtney CL, Rapoport B & McLachlan SM Thyroid autoantibodies are rare in nonhuman great apes and hypothyroidism cannot be attributed to thyroid autoimmunity. Endocrinology 154, 4896–4907 (2013). [PubMed: 24092641]

13. Sun GH, DeMonner S & Davis MM Epidemiological and economic trends in inpatient and outpatient thyroidectomy in the United States, 1996–2006. Thyroid 23, 727–733 (2013). [PubMed: 23173840]

14. Stergachis AB, Debo BM, Haugen E, Churchman LS & Stamatoyannopoulos JA Single-molecule regulatory architectures captured by chromatin fiber sequencing. Science 368, 1449–1454 (2020). [PubMed: 32587015]

15. Jha A et al. DNA-m6A calling and integrated long-read epigenetic and genetic analysis with fibertools. Preprint at bioRxiv 10.1101/2023.04.20.537673 (2023).

16. Dubocanin D et al. Conservation of chromatin organization within human and primate centromeres. Preprint at bioRxiv 10.1101/2023.04.20.537689 (2023).

17. Koren S et al. De novo assembly of haplotype-resolved genomes with trio binning. Nat. Biotechnol 36, 1174–1182 (2018).

18. Zannini M et al. TTF-2, a new forkhead protein, shows a temporal expression in the developing thyroid which is consistent with a role in controlling the onset of differentiation. EMBO J 16, 3185–3197 (1997). [PubMed: 9214635]

19. Cuesta I, Zaret KS & Santisteban P The forkhead factor FoxE1 binds to the thyroperoxidase promoter during thyroid cell differentiation and modifies compacted chromatin structure. Mol. Cell. Biol 27, 7302–7314 (2007). [PubMed: 17709379]

20. Zaret KS & Carroll JS Pioneer transcription factors: establishing competence for gene expression. Genes Dev 25, 2227–2241 (2011). [PubMed: 22056668]

21. Kalinowski FC et al. microRNA-7: a tumor suppressor miRNA with therapeutic potential. Int. J. Biochem. Cell Biol 54, 312–317 (2014). [PubMed: 24907395]

22. Kefas B et al. microRNA-7 inhibits the epidermal growth factor receptor and the Akt pathway and is down-regulated in glioblastoma. Cancer Res 68, 3566–3572 (2008). [PubMed: 18483236]

23. Webster RJ et al. Regulation of epidermal growth factor receptor signaling in human cancer cells by microRNA-7. J. Biol. Chem 284, 5731–5741 (2009). [PubMed: 19073608]

24. Jiang L et al. MicroRNA-7 targets IGF1R (insulin-like growth factor-1 receptor) in tongue squamous cell carcinoma cells. Biochem. J 432, 199–205 (2010). [PubMed: 20819078]

25. Zhao X et al. MicroRNA-7 functions as an anti-metastatic microRNA in gastric cancer by targeting insulin-like growth factor-1 receptor. Oncogene 32, 1363–1372 (2013). [PubMed: 22614005]

26. Fernandez-de Frutos M et al. MicroRNA 7 impairs insulin signaling and regulates aβ levels through posttranscriptional regulation of the insulin receptor substrate 2, insulin receptor, insulin-degrading enzyme, and liver X receptor pathway. Mol. Cell. Biol 39, e00170–19 (2019). [PubMed: 31501273]

27. Fang Y, Xue JL, Shen Q, Chen J & Tian L MicroRNA-7 inhibits tumor growth and metastasis by targeting the phosphoinositide 3-kinase/Akt pathway in hepatocellular carcinoma. Hepatology 55, 1852–1862 (2012). [PubMed: 22234835]

28. Wang Y, Liu J, Liu C, Naji A & Stoffers DA MicroRNA-7 regulates the mTOR pathway and proliferation in adult pancreatic β-cells. Diabetes 62, 887–895 (2013). [PubMed: 23223022]

29. Augenlicht A et al. MiR-7–5p inhibits thyroid cell proliferation by targeting the EGFR/MAPK and IRS2/PI3K signaling pathways. Oncotarget 12, 1587–1599 (2021). [PubMed: 34381564]

30. Brewer C, Yeager N & Di Cristofano A Thyroid-stimulating hormone initiated proliferative signals converge in vivo on the mTOR kinase without activating AKT. Cancer Res 67, 8002–8006 (2007). [PubMed: 17804710]

31. Coulonval K et al. Phosphatidylinositol 3-kinase, protein kinase B and ribosomal S6 kinases in the stimulation of thyroid epithelial cell proliferation by cAMP and growth factors in the presence of insulin. Biochem. J 348, 351–358 (2000). [PubMed: 10816429]

32. Chen M, Chen LM, Lin CY & Chai KX Hepsin activates prostasin and cleaves the extracellular domain of the epidermal growth factor receptor. Mol. Cell. Biochem 337, 259–266 (2010). [PubMed: 19911255]

33. Ding M, Bruick RK & Yu Y Secreted IGFBP5 mediates mTORCI-dependent feedback inhibition of IGF-1 signalling. Nat. Cell Biol 18, 319–327 (2016). [PubMed: 26854565]

34. Katz M et al. A reciprocal tensin-3-cten switch mediates EGF-driven mammary cell migration. Nat. Cell Biol 9, 961–969 (2007). [PubMed: 17643115]

35. Jia S et al. Essential roles of PI(3)K-p110β in cell growth, metabolism and tumorigenesis. Nature 454, 776–779 (2008). [PubMed: 18594509]

36. Porcu E et al. A meta-analysis of thyroid-related traits reveals novel loci and gender-specific differences in the regulation of thyroid function. PLoS Genet 9, e1003266 (2013).

37. Malinowski JR et al. Genetic variants associated with serum thyroid stimulating hormone (TSH) levels in European Americans and African Americans from the eMERGE Network. PLoS ONE 9, e111301 (2014).

38. Rodriguez CM & Todd PK New pathologic mechanisms in nucleotide repeat expansion disorders. Neurobiol. Dis 130, 104515 (2019).

39. Steely CJ, Watkins WS, Baird L & Jorde LB The mutational dynamics of short tandem repeats in large, multigenerational families. Genome Biol 23, 253 (2022). [PubMed: 36510265]

40. Chen LL & Yang L ALUternative regulation for gene expression. Trends Cell Biol 27, 480–490 (2017). [PubMed: 28209295]

41. Su M, Han D, Boyd-Kirkup J, Yu X & Han JJ Evolution of Alu elements toward enhancers. Cell Rep 7, 376–385 (2014). [PubMed: 24703844]

42. Maston GA & Ruvolo M Chorionic gonadotropin has a recent origin within primates and an evolutionary history of selection. Mol. Biol. Evol 19, 320–335 (2002). [PubMed: 11861891]

43. Szkudlinski MW, Teh NG, Grossmann M, Tropea JE & Weintraub BD Engineering human glycoprotein hormone superactive analogues. Nat. Biotechnol 14, 1257–1263 (1996). [PubMed: 9631089]

44. Glinoer D The regulation of thyroid function in pregnancy: pathways of endocrine adaptation from physiology to pathology. Endocr. Rev 18, 404–433 (1997). [PubMed: 9183570]

45. Ock S et al. Thyrocyte-specific deletion of insulin and IGF-1 receptors induces papillary thyroid carcinoma-like lesions through EGFR pathway activation. IntJ. Cancer 143, 2458–2469 (2018).

46. Bisi H et al. The prevalence of unsuspected thyroid pathology in 300 sequential autopsies, with special reference to the incidental carcinoma. Cancer 64, 1888–1893 (1989). [PubMed: 2676140]

47. Kong X et al. MicroRNA-7 inhibits epithelial-to-mesenchymal transition and metastasis of breast cancer cells via targeting FAK expression. PLoS ONE 7, e41523 (2012).

48. Romitti M et al. Transplantable human thyroid organoids generated from embryonic stem cells to rescue hypothyroidism. Nat. Commun 13, 7057 (2022). [PubMed: 36396935]

49. Ikegami K, Refetoff S, Van Cauter E & Yoshimura T Interconnection between circadian clocks and thyroid function. Nat. Rev. Endocrinol 15, 590–600 (2019). [PubMed: 31406343]

50. Lem AJ et al. Serum thyroid hormone levels in healthy children from birth to adulthood and in short children born small for gestational age. J. Clin. Endocrinol. Metab 97, 3170–3178 (2012). [PubMed: 22736771]

51. Dickel DE et al. Ultraconserved enhancers are required for normal development. Cell 172, 491–499 (2018). [PubMed: 29358049]

52. Snetkova V et al. Ultraconserved enhancer function does not require perfect sequence conservation. Nat. Genet 53, 521–528 (2021). [PubMed: 33782603]

53. Villar D et al. Enhancer evolution across 20 mammalian species. Cell 160, 554–566 (2015). [PubMed: 25635462]

54. Vierstra J et al. Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. Science 346, 1007–1012 (2014). [PubMed: 25411453]

55. Franchini LF & Pollard KS Human evolution: the non-coding revolution. BMC Biol 15, 89 (2017). [PubMed: 28969617]

56. ENCODE Project Consortium et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature 583, 699–710 (2020). [PubMed: 32728249]

57. Zhang K et al. A single-cell atlas of chromatin accessibility in the human genome. Cell 184, 5985–6001 e19 (2021). [PubMed: 34774128]

58. Mencia A et al. Mutations in the seed region of human miR-96 are responsible for nonsyndromic progressive hearing loss. Nat. Genet 41, 609–613 (2009). [PubMed: 19363479]

59. Hughes AE et al. Mutation altering the miR-184 seed region causes familial keratoconus with cataract. Am. J. Hum. Genet 89, 628–633 (2011). [PubMed: 21996275]

60. Conte I et al. MiR-204 is responsible for inherited retinal dystrophy associated with ocular coloboma. Proc. Natl Acad. Sci. USA 112, E3236–E3245 (2015). [PubMed: 26056285]

61. Grigelioniene G et al. Gain-of-function mutation of microRNA-140 in human skeletal dysplasia. Nat. Med 25, 583–590 (2019). [PubMed: 30804514]

62. De Pontual L et al. Germline deletion of the miR-17 approximately 92 cluster causes skeletal and growth defects in humans. Nat. Genet 43, 1026–1030 (2011). [PubMed: 21892160]

63. McLaren W et al. The ensembl variant effect predictor. Genome Biol 17, 122 (2016). [PubMed: 27268795]

64. Paila U, Chapman BA, Kirchner R & Quinlan AR GEMINI: integrative exploration of genetic variation and genome annotations. PLoS Comput. Biol 9, e1003153 (2013).

65. Lever EG, Refetoff S, Scherberg NH & Carr K The influence of percutaneous fine needle aspiration on serum thyroglobulin. J. Clin. Endocrinol. Metab 56, 26–29 (1983). [PubMed: 6847872]

66. Robin NI, Hagen SR, Collaco F, Refetoff S & Selenkow HA Serum tests for measurement of thyroid function. Hormones 2, 266–279 (1971). [PubMed: 5006665]

67. Karvanen J The statistical basis of laboratory data normalization. Drug Inf. J 37, 101–107 (2003).

68. Kisiel MA & Klar AS Isolation and culture of human dermal fibroblasts. Methods Mol. Biol 1993, 71–78 (2019). [PubMed: 31148079]

69. Vollger MR, Clark L & DPC D fiberseq/fibertools-rs: 0.4.2 (2024–03-21). Zenodo. 10.5281/zenodo.6913294 (2024).

70. Vollger MR & adrianas. mrvollger/k-mer-variant-phasing: 0.0.1. Zenodo. 10.5281/zenodo.10655527 (2024).

71. Poplin R et al. A universal SNP and small-indel variant caller using deep neural networks. Nat. Biotechnol 36, 983–987 (2018). [PubMed: 30247488]

72. Holt JM et al. HiPhase: jointly phasing small and structural variants from HiFi sequencing. Preprint at bioRxiv 10.1101/2023.05.03.539241v1 (2023).

73. Nurk S et al. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. Genome Res 30, 1291–1305 (2020). [PubMed: 32801147]

74. Vollger MR, Neph S & Bohaczuk S fiberseq/FIRE: v0.0.4 Fix missing fibers in the FDR calculation. Zenodo. 10.5281/zenodo.10023811 (2024).

75. Fondrie WE & Noble WS Machine learning strategy that leverages large data sets to boost statistical power in small-scale experiments. J. Proteome Res 19, 1267–1274 (2020). [PubMed: 32009418]

76. Kall L, Canterbury JD, Weston J, Noble WS & MacCoss MJ Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nat. Methods 4, 923–925 (2007). [PubMed: 17952086]

77. Langmead B, Trapnell C, Pop M & Salzberg SL Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10, R25 (2009). [PubMed: 19261174]

78. Trapnell C, Pachter L & Salzberg SL TopHat: discovering splice junctions with RNA-seq. Bioinformatics 25, 1105–1111 (2009). [PubMed: 19289445]

79. Trapnell C et al. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat. Biotechnol 31, 46–53 (2013). [PubMed: 23222703]

80. Sun Z et al. CAP-miRSeq: a comprehensive analysis pipeline for microRNA sequencing data. BMC Genomics 15, 423 (2014). [PubMed: 24894665]

81. Robinson JT et al. Integrative genomics viewer. Nat. Biotechnol 29, 24–26 (2011). [PubMed: 21221095]

82. Danecek P et al. Twelve years of SAMtools and BCFtools. GigaSciensce 10, giab008 (2021).

83. Grasberger H et al. DUOX2 variants associate with preclinical disturbances in microbiota-immune homeostasis and increased inflammatory bowel disease risk. J. Clin. Invest 131, e141676 (2021).

84. Sanchez-Navarro I et al. Comparison of gene expression profiling by reverse transcription quantitative PCR between fresh frozen and formalin-fixed, paraffin-embedded breast cancer tissues. Biotechniques 48, 389–397 (2010). [PubMed: 20569212]

85. Hsiao LL et al. A compendium of gene expression in normal human tissues. Physiol. Genomics 7, 97–104 (2001). [PubMed: 11773596]

86. Chen C et al. Real-time quantification of microRNAs by stem-loop RT-PCR. Nucleic Acids Res 33, e179 (2005). [PubMed: 16314309]

87. Mootha VK et al. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat. Genet 34, 267–273 (2003). [PubMed: 12808457]

88. Chen Y & Wang X miRDB: an online database for prediction of functional microRNA targets. Nucleic Acids Res 48, D127–D131 (2020). [PubMed: 31504780]

89. Ahsan S & Draghici S Identifying significantly impacted pathways and putative mechanisms with iPathwayGuide. Curr. Protoc. Bioinformatics 57, 7.15.1–7.15.30 (2017).

90. Volloch V, Schweitzer B & Rits S Ligation-mediated amplification of RNA from murine erythroid cells reveals a novel class of β globin mRNA with an extended 5’-untranslated region. Nucleic Acids Res 22, 2507–2511 (1994). [PubMed: 8041612]

91. Maruyama K & Sugano S Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. Gene 138, 171–174 (1994). [PubMed: 8125298]

92. Rao AR & Nelson SF Calculating the statistical significance of rare variants causal for Mendelian and complex disorders. BMC Med. Genomics 11, 53 (2018).

93. Krietenstein N et al. Ultrastructural details of mammalian chromosome architecture. Mol. Cell 78, 554–565 (2020). [PubMed: 32213324]

94. Paraskevopoulou MD et al. DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows. Nucleic Acids Res 41, W169–W173 (2013). [PubMed: 23680784]

95. Vollger MR StergachisLab/Fiber-seq-figures-for-RTSH: 0.0.1. Zenodo. 10.5281/zenodo.10655305 (2024).

96. Jurka J & Smith T A fundamental division in the Alu family of repeated sequences. Proc. Natl Acad. Sci. USA 85, 4775–4778 (1988). [PubMed: 3387438]

97. Kapitonov V & Jurka J The age of Alu subfamilies. J. Mol. Evol 42, 59–65 (1996). [PubMed: 8576965]

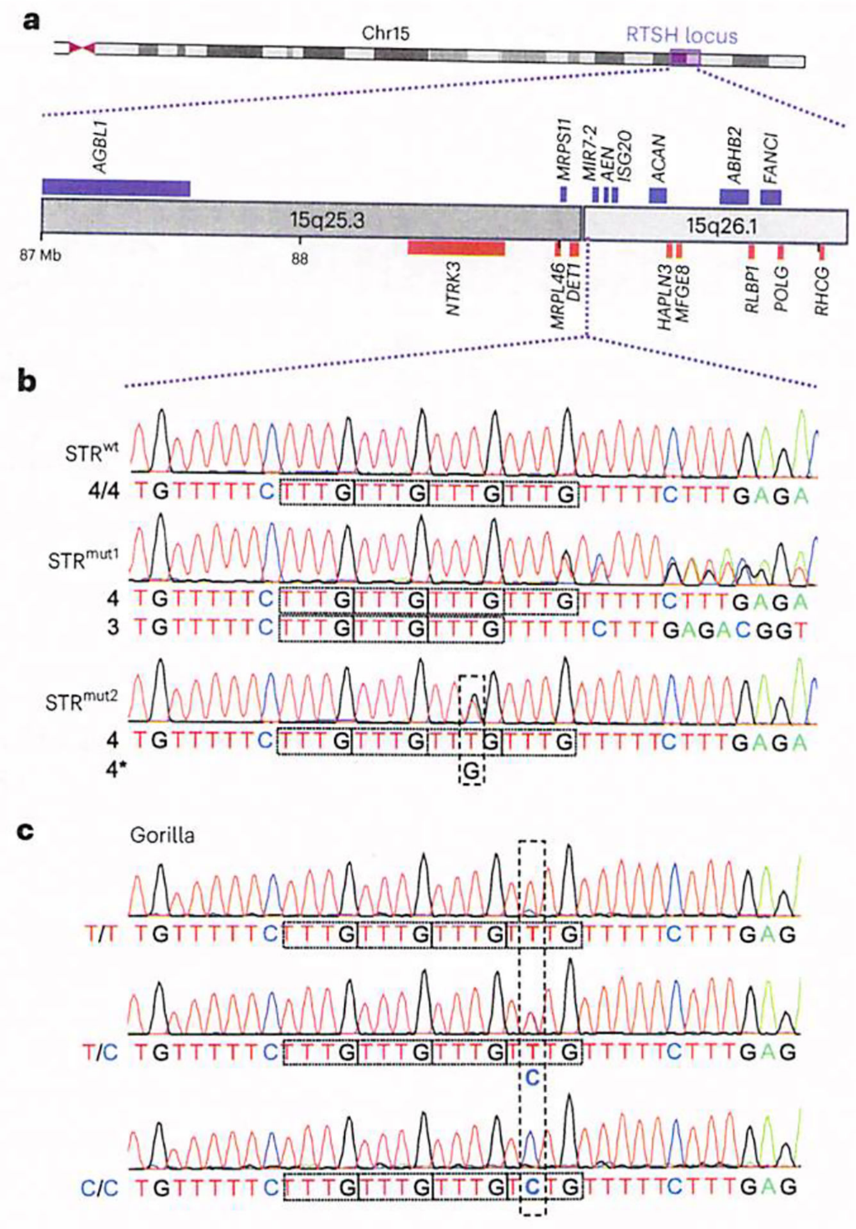


**Fig. 1|. STR variants in a primate-specific region of chr15.**

**a,** Overview of the autosomal dominant RTSH locus (OMIM: 609893) mapped in five large pedigrees[10]. **b,** Representative sequencing electropherograms of the WT STR 4/4, the heterozygous 3 repeat variant (STR 4/3) and the heterozygous T > G SNV in the third repeat (STR 4/4*). **c,** Sequences of the STR variants in *Gorilla,* homozygous $(TTTG)_4$, and heterozygous and homozygous replacement of the second T with a C in the fourth repeat.

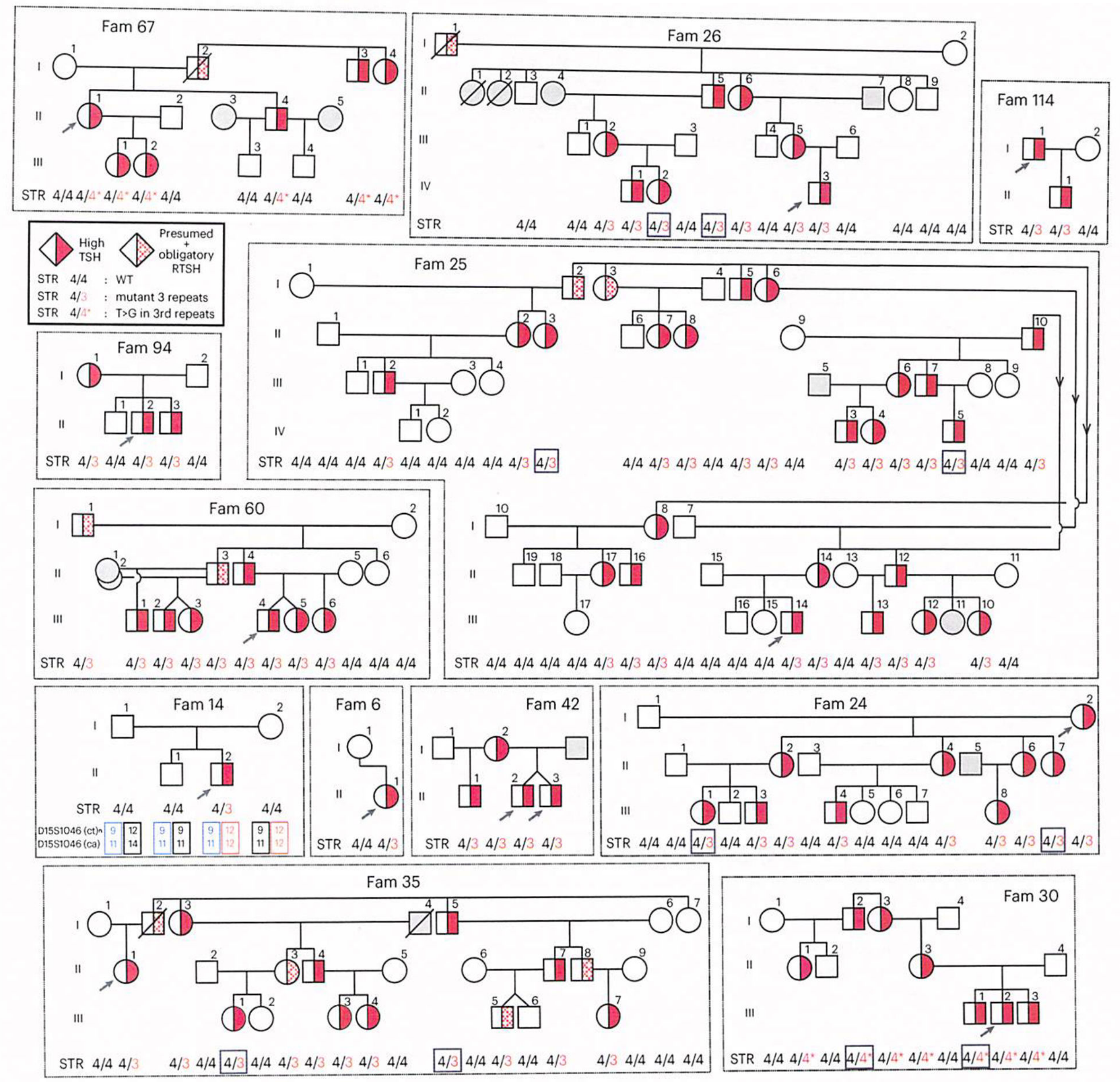

Fam 67
STR 4/4 4/4* 4/4* 4/4* 4/4 4/4 4/4* 4/4 4/4* 4/4*
Fam 26
STR 4/4 4/4 4/3 4/3 4/3 4/4 4/3 4/3 4/4 4/3 4/3 4/4 4/4 4/4 4/4
Fam 114
STR 4/3 4/3 4/4
High TSH
Presumed + obligatory RTSH
STR 4/4 : WT
STR 4/3 : mutant 3 repeats
STR 4/4* : T>G in 3rd repeats
Fam 94
STR 4/3 4/4 4/3 4/3 4/4
Fam 25
STR 4/4 4/4 4/4 4/3 4/4 4/4 4/4 4/4 4/3 4/3 4/4 4/3 4/3 4/4 4/3 4/3 4/4 4/3 4/3 4/3 4/3 4/3 4/4 4/4 4/3
STR 4/4 4/4 4/4 4/4 4/3 4/3 4/3 4/4 4/4 4/4 4/4 4/3 4/3 4/4 4/3 4/3 4/3 4/3 4/4
Fam 60
STR 4/3 4/3 4/3 4/3 4/3 4/3 4/3 4/3 4/3 4/4 4/4 4/4
Fam 14
STR 4/4 4/4 4/3 4/4
D15S1046 (ct)n 9 12 9 9 9 12 9 12
D15S1046 (ca) 11 14 11 11 11 12 11 12
Fam 6
STR 4/4 4/3
Fam 42
STR 4/3 4/3 4/3 4/3
Fam 24
STR 4/4 4/4 4/3 4/4 4/3 4/3 4/4 4/3 4/4 4/4 4/4 4/3 4/3 4/3 4/3 4/3
Fam 35
STR 4/4 4/3 4/3 4/4 4/3 4/4 4/3 4/3 4/3 4/3 4/4 4/3 4/4 4/3 4/4 4/3 4/3 4/4 4/4 4/4
Fam 30
STR 4/4 4/4* 4/4 4/4* 4/4* 4/4* 4/4 4/4* 4/4* 4/4* 4/4

**Fig. 2|. Pedigrees of families with STR mutations showing genotype-phenotype correlation.**

The RTSH phenotype is indicated by the half-colored area in each symbol and the genotype aligned below each symbol. Genotypes in boxes correspond to participants screened by WGS. Of note, the haplotypes for family 14 indicate a de novo mutation.

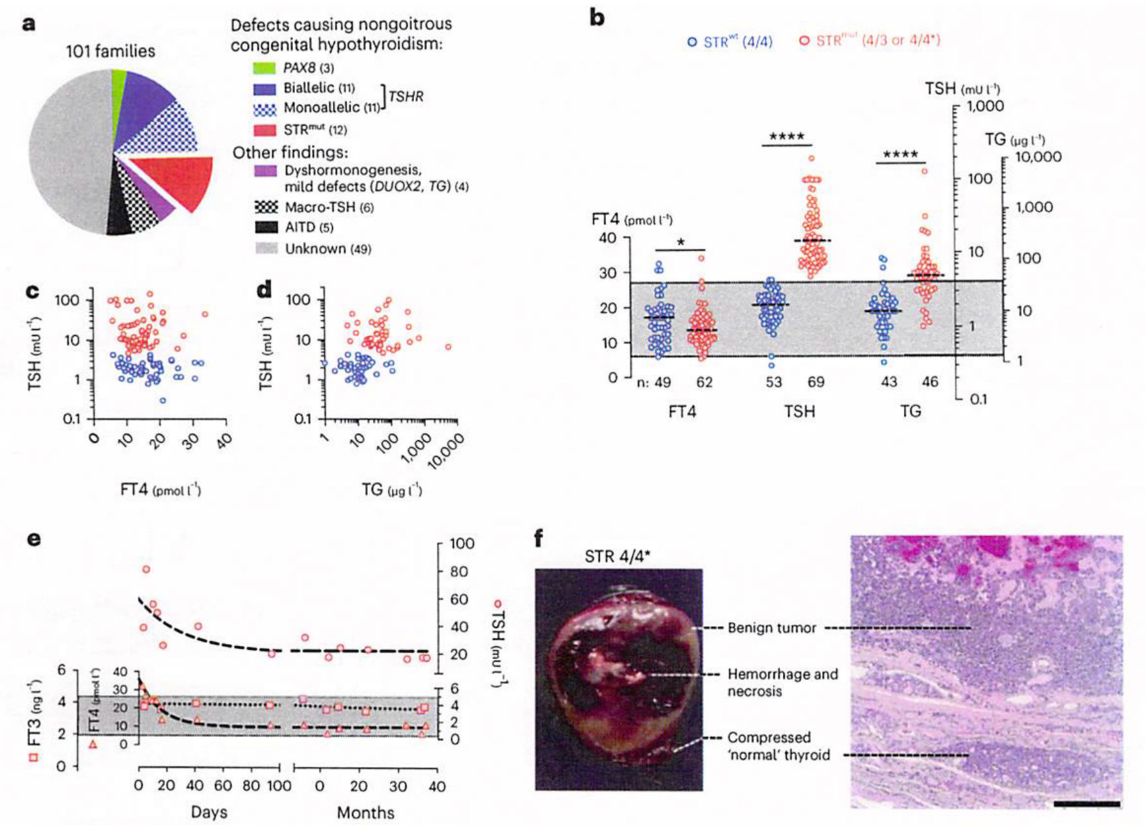


**Fig. 3|. Etiology of familial RTSH and thyroid phenotype in individuals with STR$^{mut}$.**

**a,** Findings in 101 RTSH families. All the probands had compensated RTSH. The number of families is indicated in parenthesis. **b,** FT4, TSH and TG concentr ations in serum of individuals with STR$^{mut}$ and their unaffected relatives. For participants on L-T4, treatment was discontinued for at least 6 weeks prior to thyroid evaluation. Note that scales were adjusted for the respective tests to produce a common reference range (shaded in gray). The TG value of 5,500 and one of 700 μg l$^{-1}$ are from participants II-4 of family 60 and I-2 of family 30, both of whom developed goiters that were removed surgically. *$P = 0.026$; ****$p < 0.0001$ (two-tailed Mann-Whitney). **c,** Scatterplot of TSH versus FT4 for STR$^{wt}$ ($n = 49$; blue) and STR$^{mut}$ ($n = 61$; red). **d,** Scatterplot of TSH versus TG for STR$^{wt}$ ($n = 43$; blue) and STR$^{mut}$ ($n = 45$; red). **e,** The evolution of thyroid tests from birth in participant II-3 of family 94 (STR 4/3). Values follow the expected age-related changes except for the higher concentrations of TSH. **f,** Macroscopic and histological appearance of the excised thyroid gland of participant I-2 of family 30. Scale bar = 500 μm.

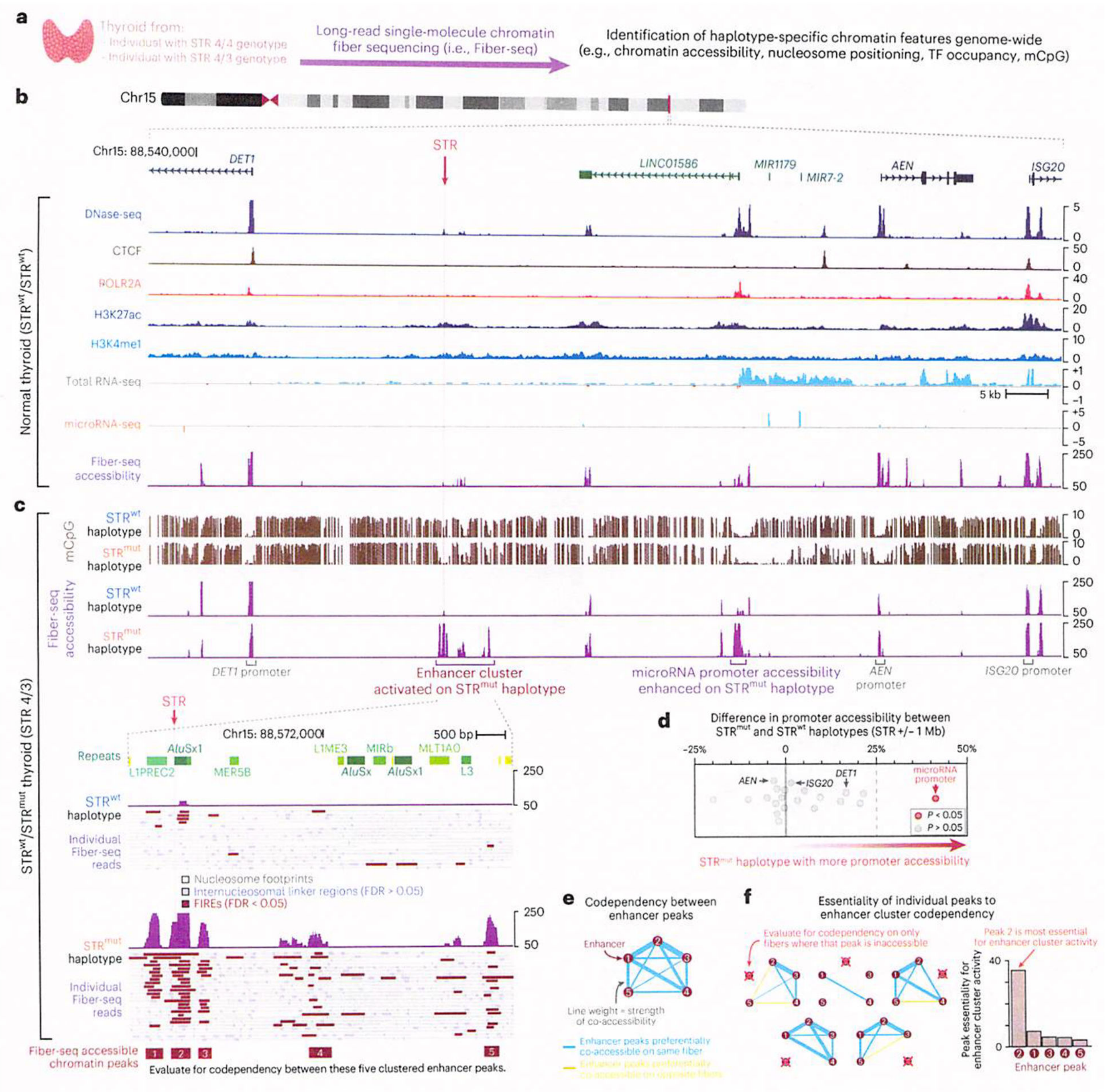


**Fig. 4 |. STR$^{mut}$ is associated with the activation of a codependent thyroid-specific enhancer cluster.**

**a,** Fiber-seq was performed on thyroid tissue from a healthy control individual, as well as an individual with RTSH heterozygous for STR$^{mut}$ to obtain haploty pe-resolved chromatin and CpG methylation pattern. **b,** Genomic locus showing chromatin and RNA expression data from a healthy individual, STR$^{wt}$. DNase-seq, ChIP-seq and RNA-seq data were obtained from ENCODE. **c,** Genomic locus showing Fiber-seq-derived chromatin data from an individual with RTSH (STR 4/3). Bottom left: genomic locus showing single-molecule chromatin architectures derived from Fiber-seq, with nucleosomes, internucleosomal linker regions and FIREs colored in gray, lavender and red, respectively. Aggregate single-molecule accessibility patterns are displayed above. **d,** Swarm plot showing the difference in promoter accessibility between he STR$^{mut}$ and STR$^{wt}$ haplotypes for all genes within 1 Mb of the STR. *P* values were calculated using a two-sided Fisher's exact test. **e,** Network analysis showing the pairwise codependency between each of the five accessible elements shown in **c.** Blue lines indicate peaks that are preferentially accessible along the same molecule, and the thick lines indicate the strength of codependency. **f,** Essentiality analysis for each of the five regulatory elements within this codependent enhancer cluster. Left: five separate networks showing the pairwise codependency after removal of reads with an accessible FIRE overlapping the element with a red cross symbol over it. Right: bar plot showing the difference in codependency of this enhancer cluster after removal of each of the five elements. FIRE, Fiber-seq inferred regulatory element.

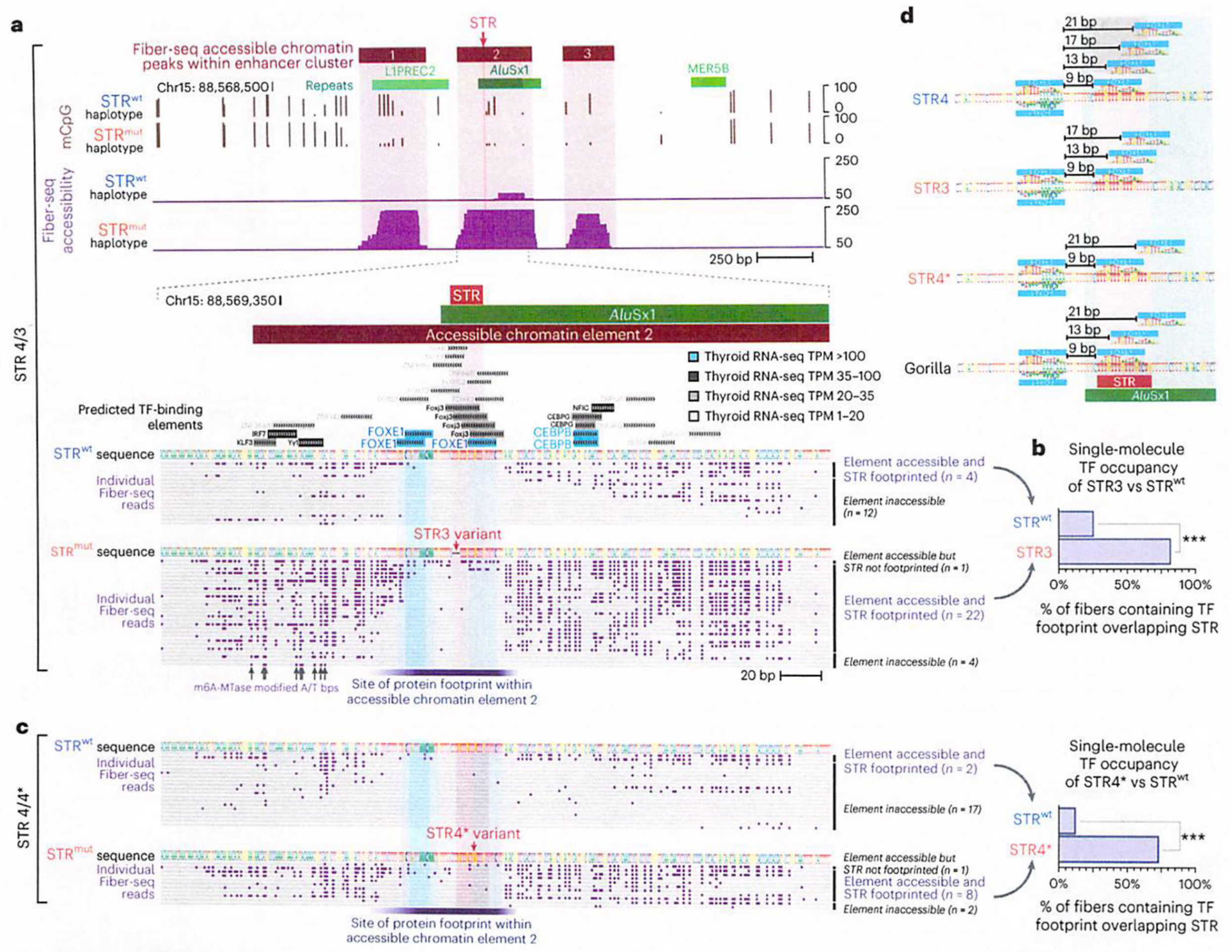


**Fig.5|. $\mathbf{STR^{mut}}$ stabilizes TF-binding occupancy.**

**a,** Single-molecule protein occupancy overlapping the $STR^{wt}$ and $STR^{mut}$ variant using Fiber-seq from thyroid tissue from an individual with RTSH (same as Fig. 4c). Predicted TF-binding elements using FIMO scans of JASPAR CORE elements are shown, with the GTEx RNA-seq expression level in the thyroid of each TF indicated by the color of the element (predicted binding elements for TFs with less than one TPM in thyroid were removed). m6A-MTase-modified bases along each fiber are indicated by purple marks. The region corresponding to a large protein footprint overlapping the $STR^{wt}$ and $STR^{mut}$ variants is indicated in blue. **b,** Bar plot showing the proportion of fibers along the $STR^{wt}$ and $STR^{mut}$ haplotype that demonstrate a protein occupancy event overlapping the STR site. ***$P = 0.00039$; two-sided Fisher's exact test. **c,** Fiber-seq was performed on thyroid tissue from an individual with RTSH heterozygous for the STR4* variant. Shown are m6A-MTase-modified bases along each fiber from the STR4 and STR4* haplotypes, as well as a bar plot comparing protein occupancy at the STR site. ***$P = 0.00097$; two-sided Fisher's exact test. **d,** Sequences of the STR4, STR3, STR4* and *Gorilla* $(TTTG)_3(TCTG)_1$ alleles, as well as the predicted location of FOXE1 binding elements along these alleles. TPM, transcript per million.

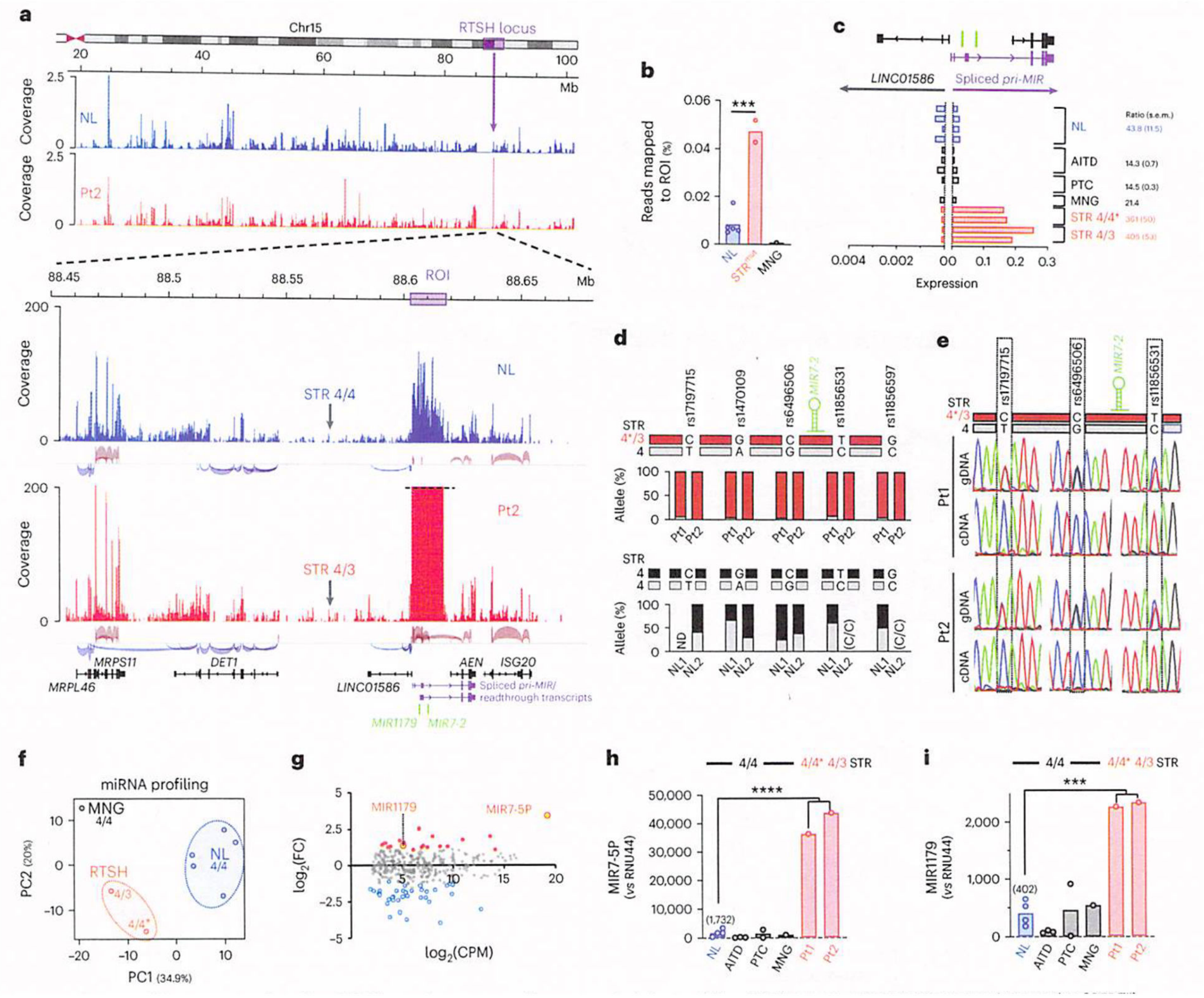


**Fig. 6|. Allele-specific overexpression of a miRNA locus downstream of STR$^{mut}$ in the participants' thyroid glands.**

**a,** Mapping of RNA-seq paired-end reads identifies a ~13.5 kb ROI35 kb q-terminal of STR$^{mut}$ that comprises*MIR7–2* and *MIR1179* and shows exceptionally high transcriptional activity in STR$^{mut}$ carriers. Boundaries of the ROI coincide with the major transcription start site (Supplementary Fig. 4) and a polyA site (chr15:88, 604, 683–88, 618, 202; hg38). **b,** Expression of the miRNA locus (ROI) in two thyroid samples representing the two different mutations (STR4/3 and 4/4*), five NL thyroids (STR$^{wt}$) and one MNG (STR$^{wt}$). ***$P =$ 0.0003; two-tailed Student's *t* test. **c,** Directional effect of STR$^{mut}$ on the expression of the miRNA locus. The relative expression of a spliced *pri-MIR* readthrough transcript and of a non-coding spliced RNA (*LINC01586)* expressed on the opposite strand from a common, bidirectionally active promoter region (Extended Data Fig. 6) is shown. The ratio values indicate the expression of readthrough transcripts versus *LINC01586*. For both Pt1 and Pt2, the data shown are from two separate specimens from their excised thyroid glands. **d,** Allelic imbalance of the miRNA locus consistent with *cis*-regulatory role of STR$^{mut}$. Allelic expression of heterozygous SNVs on the *pri-MIR* using variant calling on RNA-seq data. Segregation of the red haplotype with either STR4* (Pt1) or STR3 (Pt2) was confirmed by genotyping the respective pedigrees. **e,** Apparent loss-of-heterozygosity at the *pri-MIR* locus is shown by amplification from $STR^{mut}$ thyroid cDNA. **f,** Mature miRNA profile in $STR^{mut}$ thyroids. PCA distinguishes $STR^{mut}$ from $STR^{wt}$ glands (either NL or MNG). **g,** MA plot of thyroidal miRNA expression *(n* = 2 $STR^{mut}$ versus *n* = 5 NL with $STR^{wt}$). MiRNAs substantially upregulated or downregulated in $STR^{mut}$ (FDR < 0.05; $|\log_2(FC)| > 0.58$) are indicated by red or blue dots, respectively. CPM, TMM normalized counts per million of mature miRNAs. **h** and **i,** Real-time PCR assays for MIR7–5P and MIR1179. Data points represent individuals with NL *(n* = 4), AITD *(n* = 3), PTC *(n = 2),* MNG (*n* = 1) or $STR^{mut}$ (*n* = 2; mean values from two separate specimens). ***$p = 0.0003$; ***$p < 0.0001$; two-tailed Student's *t* test.

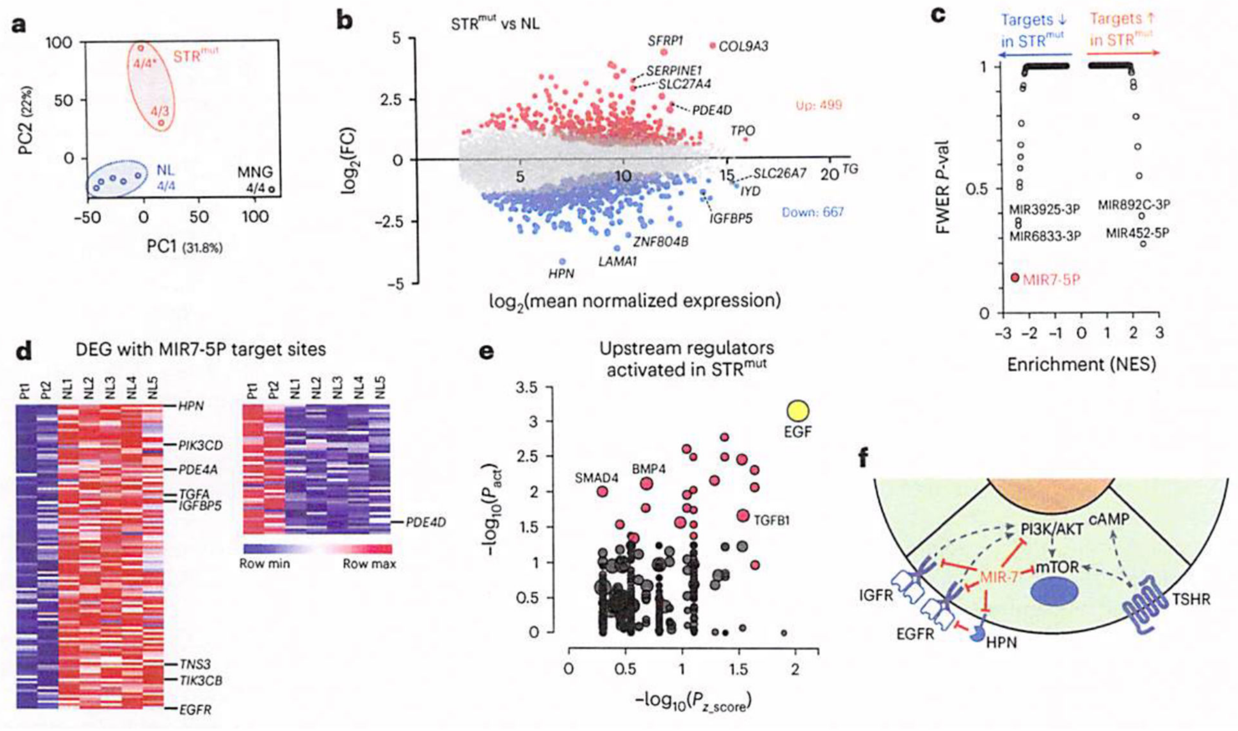


**Fig. 7|. Overexpression of MIR7–5P in STR$^{mut}$ thyroids is linked to altered proliferative signaling pathways.**

**a,** PCA plot of thyroidal mRNA expression determined by paired-end RNA-seq. **b,** MA plot indicating DEGs (FDR < 0.05 with $|\log_2(FC)| > 0.58$; STR$^{mut}$ versus NL) by red (higher in STR$^{mut}$) or blue (lower in STR$^{muI}$) dots. **c.** Evidence for higher MIR7–5P activity in STR$^{mut}$ thyroid glands. A total of 2,141 miRNA target gene sets (miRDB subset from MSigDB) were tested for their correlation with gene expression changes (STR$^{mut}$ versus NL) using gene set enrichment analysis. The MIR7–5P target gene set was found to have the most significant enrichment for genes downregulated in STR$^{mut}$. **d,** Expression heatmaps of all DEG identified as MIR7–5P targets in MicroT-CDS (miTG score of >0.6). **e,** Prediction of EGF as activated upstream regulator (FDR = 0.059) by analyzing the differential expression of downstream genes (iPathwayGuide). The *x* axis position is the log of the unadjusted *P* value based on the activation *z* score, which is derived by comparison with a model that assigns random regulation. The *y* axis position is the log of the unadjusted overrepresentation *P* value based on the number of DE target genes consistent with the activation profile. The size of each dot represents the number of DEG directionally consistent with the regulator profile. **f,** Proposed working model for thyroid pathophysiology in STR$^{mut}$. Abnormally high MIR7–5P level in thyrocytes impairs proliferative response compatible with the phenotype of non-goitrous congenital hyperthyrotropinemia. The predicted activation of EGF downstream signaling conceivably would promote proliferative lesions observed in a subset of adult participants. Potential mechanisms are an increased EGF tone because of an abnormal feedback response to suppressed IGFR signaling[45] or reduced EGFR inactivation by HPN[32]. FWER, family-wise error rate; NES, normalized enrichment score.

## Table 1|

variants causing RTSH at the chr15 locus, families and annotations for each variant

| Families with variant | Variant in VCF and HGVS format (GRCh37, GRCh38) | Label used | Frequency in population controls | | CADD phred (v1.6) | gnomAD genomic constraint (z score, percentile) | SCREEN registry | ClinGen allele ID |
|---|---|---|---|---|---|---|---|---|
| | | | gnomAD v3.1.2 | gnomAD v2.1,1 | | | | |
| 6, 14, 24, 25, 26, 35, 42, 60, 94, 114 | GRCh37:15:89112664:CTTTG:C<br>GRCh38:l 5:88569433:CTTTG:C<br>NC_000015.9:g.89112680_89112683del<br>NC_000015.10:g.88569449_88569452del | 4/3 | 1/151588 | 0 | 1.58 | 2.49 (10%) | Overlaps EH38E31S3006 predicted enhancer-like signature in thyroid | CA274493169 |
| 30, 67 | GRCh37:15:89112675:T:G<br>GRCh38:15:88569444:T:G<br>NC_000015,9:g.89112675T>G<br>NC_000015.10:g.88569444 T >G | 4/4* | 0 | 0 | 4.963 | 2.49 (10%) | Overlaps EH38E3153006 predicted enhancer-like signature in thyroid | CA274493177 |

‘Label used’ lndicates the shorthand label used to refer to each variant in the paper. gnomAD genomic constraint is provided as measured by z score and percentile for the flanking 1kb region of the genome (for example, 10% = top 10 percentile of constrained genes). Other very rare variants from gnomAD, not found in our study cohort. are shown in Extended Data Table 1.