Supplementary Information

Discovery of common variants associated with low TSH levels and thyroid cancer risk

Julius Gudmundsson[1*], Patrick Sulem[1*], Daniel F. Gudbjartsson[1], Jon G. Jonasson[2,3,4], Gisli Masson[1], Huiling He[5], Aslaug Jonasdottir[1], Asgeir Sigurdsson[1], Simon N. Stacey[1], Hrefna Johannsdottir[1], Hafdis Th. Helgadottir[1], Wei Li[5], Rebecca Nagy[5], Matthew D. Ringel[6], Richard T. Kloos[6], Marieke C.H. de Visser[7], Theo S. Plantinga[8], Martin den Heijer[7,15], Esperanza Aguillo[9], Angeles Panadero[10], Enrique Prats[11], Almudena Garcia[12], Ana De Juan[12], Fernando Rivera[12], G. Bragi Walters[1], Hjordis Bjarnason[1], Laufey Tryggvadottir[3,4], Gudmundur I. Eyjolfsson[13], Unnur S. Bjornsdottir[3], Hilma Holm[1] , Isleifur Olafsson[2], Kristleifur Kristjansson[1], Hoskuldur Kristvinsson[2], Olafur Th. Magnusson[1], Gudmar Thorleifsson[1], Jeffrey R. Gulcher[1], Augustine Kong[1], Lambertus A.L.M. Kiemeney[7], Thorvaldur Jonsson[2,3], Hannes Hjartarson[2], Jose I. Mayordomo[14], Romana T. Netea-Maier[15], Albert de la Chapelle[5], Jon Hrafnkelsson[2], Unnur Thorsteinsdottir[1,3], Thorunn Rafnar[1], Kari Stefansson[1,3].

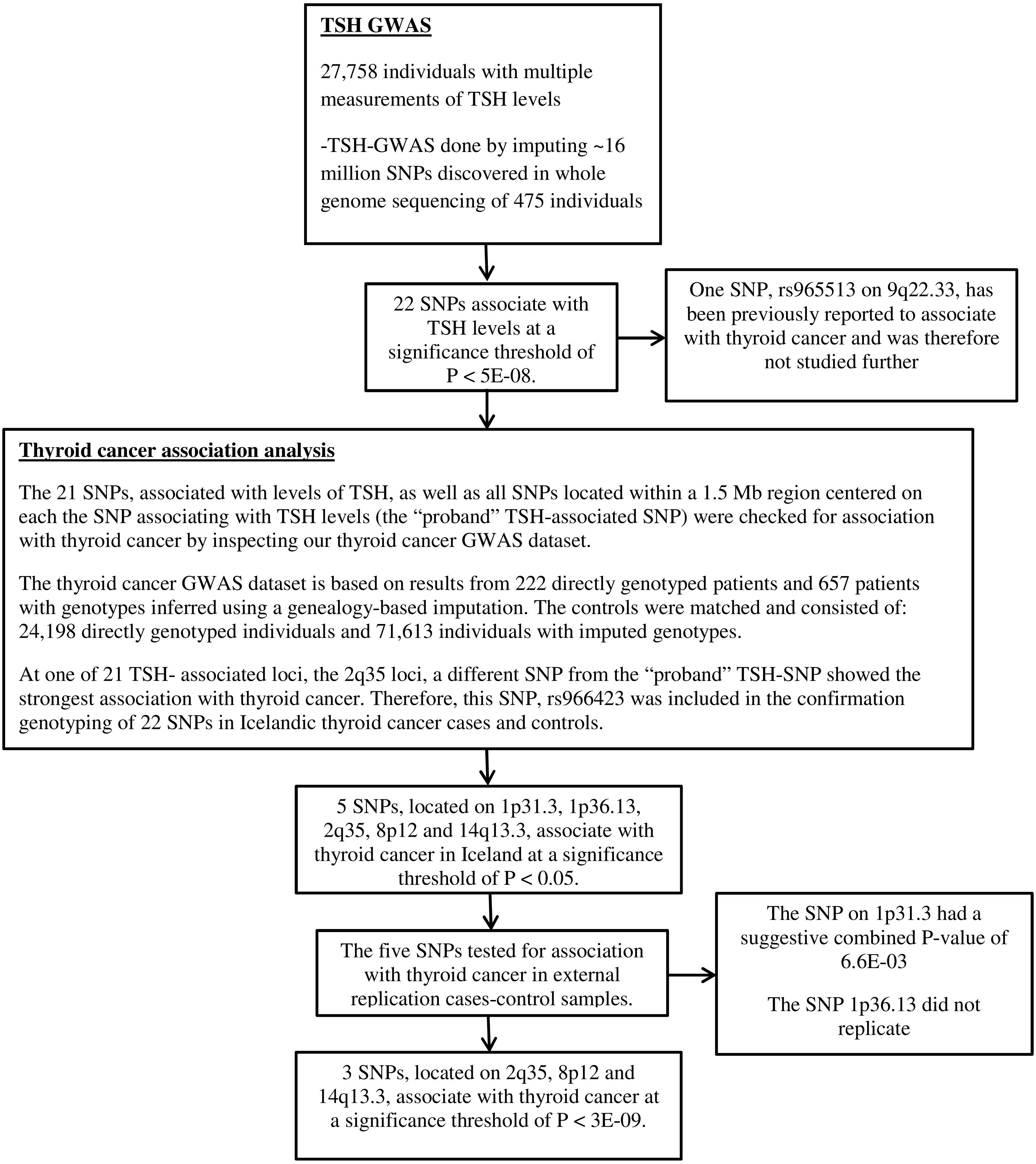
Content:

Supplementary Figures 1-3

Supplementary Tables 1-4

Supplementary Note

Nature Genetics: doi:10.1038/ng.1046



TSH GWAS

27,758 individuals with multiple measurements of TSH levels

-TSH-GWAS done by imputing ~16 million SNPs discovered in whole genome sequencing of 475 individuals

22 SNPs associate with TSH levels at a significance threshold of P < 5E-08.

One SNP, rs965513 on 9q22.33, has been previously reported to associate with thyroid cancer and was therefore not studied further

Thyroid cancer association analysis

The 21 SNPs, associated with levels of TSH, as well as all SNPs located within a 1.5 Mb region centered on each the SNP associating with TSH levels (the "proband" TSH-associated SNP) were checked for association with thyroid cancer by inspecting our thyroid cancer GWAS dataset.

The thyroid cancer GWAS dataset is based on results from 222 directly genotyped patients and 657 patients with genotypes inferred using a genealogy-based imputation. The controls were matched and consisted of: 24,198 directly genotyped individuals and 71,613 individuals with imputed genotypes.

At one of 21 TSH- associated loci, the 2q35 loci, a different SNP from the "proband" TSH-SNP showed the strongest association with thyroid cancer. Therefore, this SNP, rs966423 was included in the confirmation genotyping of 22 SNPs in Icelandic thyroid cancer cases and controls.

5 SNPs, located on 1p31.3, 1p36.13, 2q35, 8p12 and 14q13.3, associate with thyroid cancer in Iceland at a significance threshold of P < 0.05.

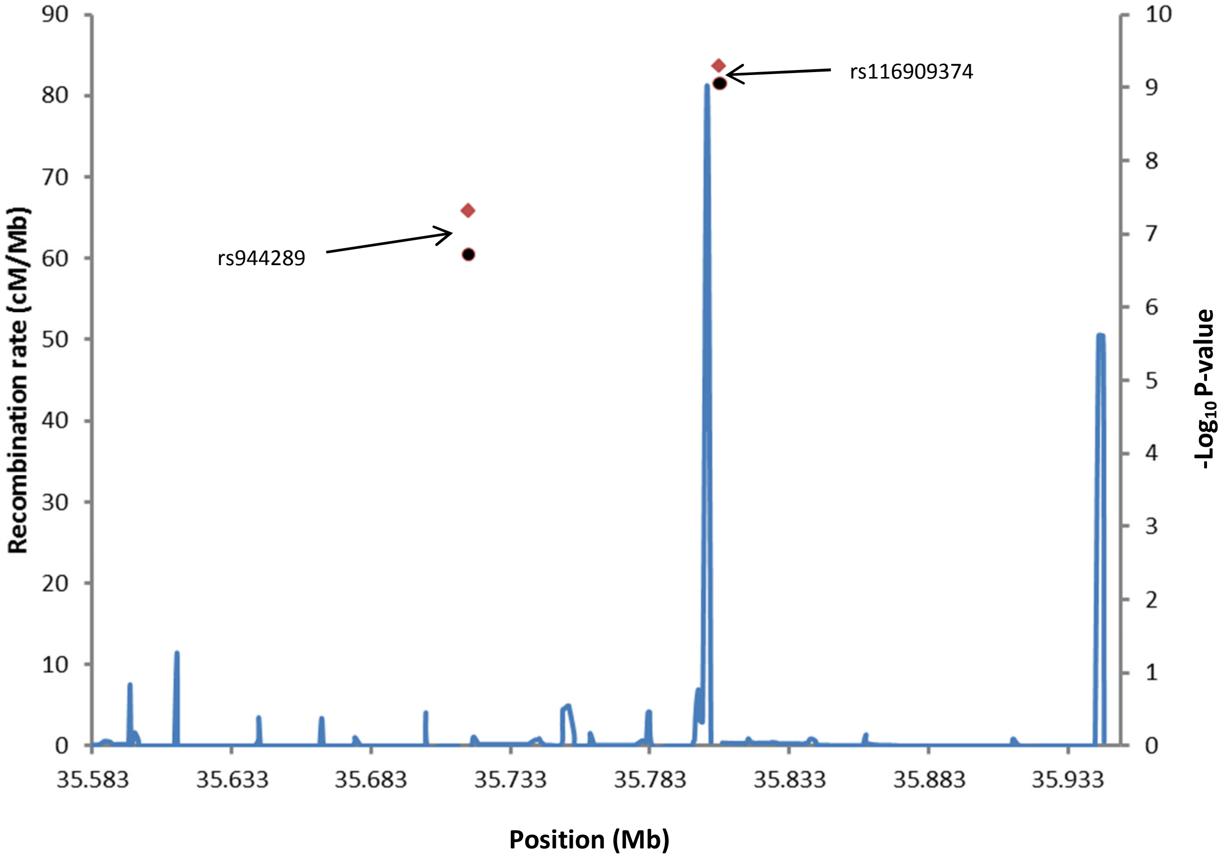
The five SNPs tested for association with thyroid cancer in external replication cases-control samples.

The SNP on 1p31.3 had a suggestive combined P-value of 6.6E-03

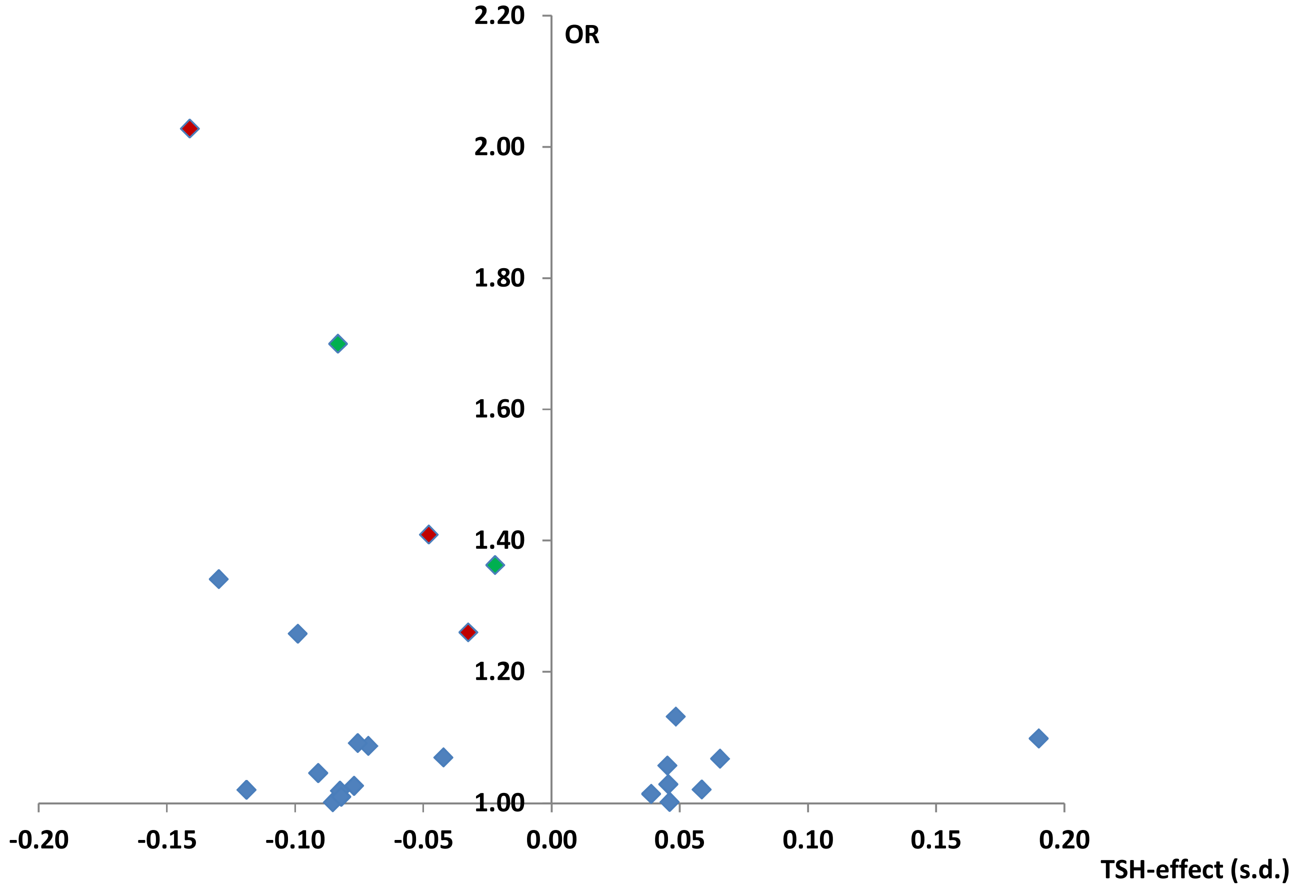
The SNP 1p36.13 did not replicate

3 SNPs, located on 2q35, 8p12 and 14q13.3, associate with thyroid cancer at a significance threshold of P < 3E-09.

Supplementary Figure 1. Summary of our study design and results



Supplementary Figure 2. Shown are the unadjusted (red diamonds) and adjusted (black circle) thyroid cancer association results ($-\log_{10}$ P-value) for rs944289 and rs116909374, as well as the recombination rate in 375 kb region on 14q13.3. The recombination rate (cM/Mb) is based on CEU HapMap phase II release 22. The association results are the combined unadjusted and adjusted results for the 4 study groups reported in Supplementary table 3.



Supplementary Figure 3. Shown is the OR for thyroid cancer (y-axis) and the TSH –effect (s.d.) (x-axis) for the 24 SNPs in Supplementary Table 2. The three red diamonds denote the three SNPs reported in Table 1 in the main text and are shown to associate with thyroid cancer at P < 3E-09. The two green diamonds denote the two previously reported thyroid cancer risk SNPs on 9q22.33 (rs965513) and 14q13.3 (rs944289). The blue diamonds denote the remaining 17 SNPs listed in Supplementary Table 2.

Supplementary Table 1a. An overview of the TSH, free-T_4 and free-T_3 measurements available for SNP-chip genotyped individuals.

Measurement type (units)	Males with measure-ments (n)	Measure-ments per male individual[a] (n)	Females with measure-ments (n)	Measure-ments per female individual[a] (n)	The median of measurements (first quartile, last quartile)
TSH (mIU/L)	10,434	4.4	17,324	6.1	1.94 (1.14, 3.16)
Free-T4 (pmol/L)	6,188	3.0	12,872	4.1	15.7 (13.8, 18.2)
Free-T3 (pmol/L)	2,981	2.3	7,044	2.8	4.6 (4.0, 5.3)

[a]The geometric mean of the number of measurements per individual

Supplementary Table 1b. The distribution of Icelandic thyroid cancer patients into different genotyping categories according to different phases of the thyroid cancer association study

i) Thyroid cancer GWAS

Sample categories	*Patients (n)*
Chip genotypes	222
Imputed genotypes	657
Neither with chip nor imputed genotypes	139
Total number of thyroid cancer patients	1,018

ii) Confirmation of thyroid cancer association by direct genotyping

Sample categories	*Patients (n)*
Chip genotyped	222
Centaurus genotyped [a]	339
Neither with chip nor Cenataurus genotypes	457
Total number of thyroid cancer patients	1,018

[a] the Centaurus genotyped samples are a subset of the 657 imputed samples in the thyroid cancer GWAS.

Supplementary Table 2. Association results in Iceland for serum levels of TSH, free-T3, free-T4, thyroid cancer, and goiter risk

			TSH levels (n = 27,758)		Thyroid cancer GWAS results (n = 222+657)[a]		Association results for directly genotyped thyroid cancer cases and controls						free T4 levels (n = 19,060)		free T3 levels (n = 10,023)		Goiter GWAS results (n = 217 +329) [a]	
SNP_Effect Allele	Chr.	Position B36 (bp)	Effect (s.d.)	P-value	OR	P-value	OR	P-value	Cases (n)	Cases (freq.)	Controls (n)	Controls (freq.)	Effect (s.d.)	P-value	Effect (s.d.)	P-value	OR	P-value
rs10799824_A	1	19,713,761	-0.099	3.8E-26	1.25	1.4E-02	1.26	0.012	524	0.183	2,330	0.151	0.030	3.7E-03	0.016	2.9E-01	1.36	2.4E-03
rs334725_C	1	61,382,637	-0.130	1.0E-20	1.27	6.8E-02	1.34	0.014	561	0.086	40,013	0.065	0.038	1.5E-02	0.023	3.0E-01	1.60	8.6E-04
rs17020124_G	1	108,159,534	-0.091	4.3E-17	1.16	1.9E-01	1.05	0.7	542	0.887	1,499	0.882	0.020	9.8E-02	0.026	1.3E-01	0.89	3.6E-01
rs11694732_C	2	1,387,357	0.045	1.6E-10	0.96	5.2E-01	1.06	0.43	545	0.415	3,179	0.401	-0.018	2.0E-02	0.012	2.9E-01	1.00	9.8E-01
rs737308_T	2	217,331,494	-0.072	5.7E-22	1.08	3.1E-01	1.09	0.26	547	0.333	3,151	0.314	0.029	3.8E-04	0.009	4.5E-01	1.08	3.4E-01
rs966423_C	2	218,018,585	-0.032	2.9E-06	1.25	1.0E-03	1.26	3.8E-04	546	0.499	38,854	0.442	0.020	9.7E-03	0.006	5.7E-01	1.08	3.5E-01
rs10030849_C	4	149,872,549	0.066	1.1E-14	1.06	5.0E-01	1.07	0.48	539	0.798	1,500	0.787	-0.001	8.8E-01	-0.012	3.6E-01	1.05	6.0E-01
rs2046045_A	5	76,571,567	-0.119	2.8E-62	0.92	2.7E-01	1.02	0.76	561	0.622	39,165	0.617	0.009	2.5E-01	0.008	4.6E-01	1.10	2.3E-01
rs729761_T	6	43,912,549	-0.085	7.5E-28	0.89	1.4E-01	1.00	0.99	543	0.263	38,820	0.263	0.021	1.7E-02	0.022	8.6E-02	1.20	4.0E-02
rs6923866_T	6	44,009,162	0.048	5.4E-09	1.01	9.3E-01	1.13	0.16	545	0.809	3,118	0.789	-0.019	4.1E-02	0.014	2.8E-01	0.76	2.4E-03
rs3008043_A	6	165,972,622	-0.076	2.0E-23	1.15	6.5E-02	1.09	0.29	536	0.314	1,476	0.296	0.012	1.7E-01	0.003	7.8E-01	1.03	7.3E-01
rs2439302_G	8	32,551,911	-0.048	7.8E-12	1.34	2.4E-05	1.41	1.3E-06	532	0.535	3,094	0.449	0.008	3.3E-01	-0.002	8.6E-01	0.99	8.9E-01
rs965513_A	9	99,595,930	-0.083	4.3E-31	1.71	6.9E-15	1.70	3.0E-18	558	0.480	43,108	0.352	-0.040	7.84E-07	0.024	4.0E-02	0.77	2.4E-03
rs7913135_C	10	101,308,629	0.046	2.8E-11	0.91	1.8E-01	1.00	0.98	502	0.484	1,494	0.484	-0.025	1.1E-03	0.006	5.8E-01	0.95	5.6E-01
rs7128207_G	11	45,186,637	-0.042	1.3E-09	1.05	5.2E-01	1.07	0.35	545	0.472	2,135	0.455	0.023	3.0E-03	0.011	3.2E-01	1.05	5.7E-01
rs61938844_A	12	95,107,989	0.190	5.6E-16	1.17	4.7E-01	1.10	0.69	544	0.028	1,610	0.025	-0.020	4.3E-01	-0.081	3.5E-02	0.98	9.4E-01
rs944289_T	14	35,718,997	-0.022	1.5E-03	1.42	5.6E-07	1.36	5.2E-07	560	0.632	39,864	0.558	0.017	2.7E-02	-0.003	7.8E-01	1.05	5.4E-01
rs116909374_T	14	35,808,112	-0.141	1.1E-16	1.61	1.5E-03	2.03	5.4E-07	542	0.085	3,190	0.044	0.041	2.8E-02	0.039	1.4E-01	1.24	2.2E-01
rs34269820_T	14	92,635,908	-0.083	7.9E-19	1.11	2.7E-01	1.02	0.85	545	0.835	1,496	0.832	0.016	1.1E-01	0.015	3.1E-01	1.25	4.4E-02
rs73362602_T	14	104,301,640	0.039	4.1E-08	0.91	2.1E-01	1.01	0.85	536	0.466	1,496	0.463	-0.003	7.5E-01	-0.021	7.0E-02	0.92	3.1E-01
rs73398284_T	15	47,501,692	0.059	4.9E-14	0.95	4.7E-01	1.02	0.81	526	0.726	1,472	0.722	-0.021	1.4E-02	-0.022	6.9E-02	0.66	1.0E-06
rs7190187_T	16	78,292,679	-0.077	3.8E-24	1.19	1.7E-02	1.03	0.75	541	0.310	1,475	0.304	0.013	1.2E-01	0.021	9.0E-02	1.09	3.1E-01
rs10420008_G	19	7,185,575	0.045	4.3E-08	0.87	9.2E-02	1.03	0.76	535	0.223	1,499	0.218	-0.009	3.4E-01	-0.026	5.6E-02	0.78	9.8E-03
rs6082762_A[b]	20	22,573,016	-0.082	3.5E-19	1.04	6.7E-01	1.01	0.85	531	0.193	39,034	0.191	0.021	4.0E-02	0.006	6.7E-01	1.35	2.0E-03

Association results are for the effect-allele of the 24 SNPs. 21 were selected based on a P-values threshold ($P < 5E-08$) for an association with serum levels of TSH, one SNP (rs966423) was selected based on its stronger two-way association with thyroid cancer, also included are the two previously published thyroid cancer risk SNPs; rs965513 on 9q22.33 and rs944289 on 14q13.3. The effect size for serum levels of TSH, free-T3 and free-T4 is measured in standard-deviation units (s.d.). The minus (-) sign in front of the effect size stands for a decreasing effect whereas no sign stands for an increasing effect. [a] The thyroid cancer GWAS results are based on 222 chip genotyped cases and 657 cases genotyped using imputation (two-way imputation) and chip genotyped 1st or 2nd degree relatives; for goiter the corresponding numbers are 217 chip genotyped patients and 329 patents with chip genotyped 1st or 2nd degree relatives. [b] The SNP rs6082762 is not on the chips used genotype the Icelandic samples and the single track assay failed in production therefore are the association results for the directly genotyped cases and controls based on data from a fully correlated SNP rs1203930 (r2 = and D' = 1 between rs6082762 and rs1203930 according to CEU HapMap data) genotyped using the Illumina chips.

Supplementary Table 3. Association results for rs334725 located on 1p31.3 and thyroid cancer in Iceland, the Netherlands, Spain and the United States

Study population (n cases/n controls)	OR	95% CI	P-value	Case (freq)	Controls (freq)
rs334725_C on 1p31.3					
Iceland (561/40,013) [a]	1.34	(1.06, 1.70)	0.014	0.086	0.065
The Netherlands (149/832)	1.22	(0.73, 2.05)	0.45	0.054	0.044
Ohio, US (357/373)	1.28	(0.77, 2.13)	0.34	0.056	0.044
Spain (89/1,399)	0.93	(0.00, ∞)	1.00	0.039	0.042
All combined (1,156/42,617) [b]	1.31	(1.08, 1.60)	6.6×10^{-3}	-	-

[a] rs334725 is present on the Illumina chips used to genotype the Icelandic GWAS population, results are included for chip-genotyped individuals. Otherwise all results for all study groups are based on single-track assay genotyping.

[b] For the combined study populations, the OR and the P value were estimated using the Mantel-Haenszel model

Supplementary Table 4. Association results for rs116909374 and rs944289 on 14q13.3, before and after adjustment

Study group	rs116909374_T		rs944289_T	
Iceland	**OR**	**P-value**	**OR**	**P-value**
Unadjusted	2.03	5.4E-07	1.36	4.2E-05
Adjusted	1.95	4.7E-07	1.30	9.6E-05
The Netherlands				
Unadjusted	1.95	0.024	1.39	0.013
Adjusted	1.93	0.028	1.38	0.014
Ohio [a]				
Unadjusted	1.60	0.26	1.51	0.0067
Adjusted	1.52	0.32	1.50	0.0078
Spain				
Unadjusted	3.37	0.0026	1.17	0.31
Adjusted	3.27	0.0040	1.13	0.45
All combined				
Unadjusted	2.07	5.0×10^{-10}	1.36	4.9×10^{-8}
Adjusted	1.99	8.7×10^{-10}	1.32	1.9×10^{-7}

Shown are results for rs116909374 before and after being adjusted for rs944289 as well as results for rs944289 before and after being adjusted for rs116909374. The two SNPs are only correlated to a very small degree (D' = 0.35 and r2 = 0.005 based on results from 3,693 Icelanders). Results are only presented for individuals where data is available for both SNPs. P_{het} is > 0.5 for both markers.

[a]For the Ohio samples data was available for both SNPs for 155 cases and 245 controls.

The LD- and correlation information the two SNPs in this table in the four different study groups is as follows:
Iceland; D’ = 0.35 r^2 = 0.0050
The Netherlands D’ = 0.13 r^2 = 0.0003
Spain; D’ = 0.63 r^2 = 0.0065
Ohio; D’ = 0.37 r^2 = 0.0026

Supplementary Note

Genotyping Methods

Illumina genotyping. The Icelandic chip-typed samples were assayed with the Illumina Human Hap300, Hap CNV370, Hap 610, 1M or Omni-1 Quad bead chips at deCODE genetics. Only the 317,503 SNPs from the Human Hap300 chip were used in the long range phasing and the subsequent SNP imputations. SNPs were excluded if they had (i) yield lower than 95%, (ii) minor allele frequency less than 1% in the population or (iii) significant deviation from Hardy-Weinberg equilibrium in the controls ($P < 0.001$), (iv) if they produced an excessive inheritance error rate (over 0.001), (v) if there was substantial difference in allele frequency between chip types (from just a single chip if that resolved all differences, but from all chips otherwise). All samples with a call rate below 97% were excluded from the analysis. The final set of SNPs used for long range phasing and GWAS was composed of 297,835 autosomal SNPs.

Single track assay SNP genotyping. Genotyping of the SNPs reported in Table 1 of the main text for the three case-control groups from Iceland, the Netherlands and Spain was carried out by deCODE Genetics in Reykjavik, Iceland, applying the Centaurus[1] (Nanogen) platform or the Illumin SNP-chips. Using the Centaurus single-track assay, we genotyped the Spanish cases and controls, the Dutch cases and controls and all the 561 Icelandic patients. Of the Icelandic patients, 222 had been previously chip genotyped for the SNPs on 1p31.3 and 2q35which are present on the Illuimina SNP-chips used in our initial GWAS genotyping effort. These 222 patients were re-genotyped using Centaurus single-track assay for confirming data consistency of the two genotyping platforms. We used Centaurus single-track assay to genotype between 1,472 and 3,190 Icelandic controls for the 21 TSH-associated SNPs. For the four TSH-associated SNPs that are present on the Illumina chips we included genotype data from 40,013 Icelandic controls GWAS study population. The 3,190 single-track assay genotyped controls are among the 40,013 Illumin chip genotyped controls and the overlap of genotype results was used to check for data consistency. Furthermore, the quality of each Centaurus SNP assay was evaluated by genotyping it in the CEU and/or YRI HapMap samples and comparing the results with the HapMap publicly released data. Assays with >1.5% mismatch rate were not used and a linkage disequilibrium (LD) test was used for markers known to be in LD.

Genotyping of samples from the Ohio study populations was done using the SNaPshot (PE Applied Biosystems,Foster City, CA) genotyping platform at the Ohio State University, as previously described[2].

*Whole Genome Sequencing***.** SNPs were imputed based on unpublished data from the Icelandic whole genomic sequencing project (457 Icelandic individuals) selected for various neoplasic, cardiovascular and psychiatric conditions. All of the individuals were sequenced to a depth of at least 10X. Sixteen million SNPs were imputed based on this set of individuals.

Sample preparation. Paired-end libraries for sequencing were prepared according to the manufacturer's instructions (Illumina). In short, approximately 5 µg of genomic DNA, isolated from frozen blood samples, was fragmented to a mean target size of 300 bp using a Covaris E210 instrument. The resulting fragmented DNA was end repaired using T4 and Klenow polymerases and T4 polynucleotide kinase with 10 mM dNTP followed by addition of an 'A' base at the ends using Klenow exo fragment (3′ to 5′-exo minus) and dATP (1 mM). Sequencing adaptors containing 'T' overhangs were ligated to the DNA products followed by agarose (2%) gel electrophoresis. Fragments of about 400 bp were isolated from the gels (QIAGEN Gel Extraction

Kit), and the adaptor-modified DNA fragments were PCR enriched for ten cycles using Phusion DNA polymerase (Finnzymes Oy) and PCR primers PE 1.0 and PE 2.0 (Illumina). Enriched libraries were further purified using agarose (2%) gel electrophoresis as described above. The quality and concentration of the libraries were assessed with the Agilent 2100 Bioanalyzer using the DNA 1000 LabChip (Agilent). Barcoded libraries were stored at −20 °C. All steps in the workflow were monitored using an in-house laboratory information management system with barcode tracking of all samples and reagents.

DNA sequencing. Template DNA fragments were hybridized to the surface of flow cells (Illumina PE flowcell, v4) and amplified to form clusters using the Illumina cBot. In brief, DNA (8–10 pM) was denatured, followed by hybridization to grafted adaptors on the flowcell. Isothermal bridge amplification using Phusion polymerase was then followed by linearization of the bridged DNA, denaturation, blocking of 3 ends and hybridization of the sequencing primer. Sequencing-by-synthesis was performed on Illumina GAIIx instruments equipped with paired-end modules. Paired-end libraries were sequenced using 2 × 101 cycles of incorporation and imaging with Illumina sequencing kits, v4. Each library or sample was initially run on a single lane for validation followed by further sequencing of ≥4 lanes with targeted cluster densities of 250–300 k/mm^2. Imaging and analysis of the data was performed using the SCS 2.6 and RTA 1.6 software packages from Illumina, respectively. Real-time analysis involved conversion of image data to base-calling in real-time.

Alignment. For each lane in the DNA sequencing output, the resulting qseq files were converted into fastq files using an in-house script. All output from sequencing was converted, and the Illumina quality filtering flag was retained in the output. The fastq files were then aligned against Build 36 of the human reference sequence using bwa version 0.5.7 (ref. [3]).

BAM file generation. SAM file output from the alignment was converted into BAM format using SAMtools version 0.1.8 (ref. [4]), and an in-house script was used to carry the Illumina quality filter flag over to the BAM file. The BAM files for each sample were then merged into a single BAM file using SAMtools. Finally, Picard version 1.17 (see http://picard.sourceforge.net/) was used to mark duplicates in the resulting sample BAM files.

*SNP calling and genotyping in whole-genome sequencing***.** A two-step approach was applied. The first step was to detect SNPs by identifying sequence positions where at least one individual could be determined to be different from the reference sequence with confidence (quality threshold of 20) based on the SNP calling feature of the pileup tool SAMtools[4]. SNPs that always differed heterozygous or homozygous from the reference were removed. The second step was to use the pileup tool to genotype the SNPs at the positions that were flagged as polymorphic. Because sequencing depth varies and hence the certainty of genotype calls also varies, genotype likelihoods rather than deterministic calls were calculated (see below). Of the 2.5 million SNPs reported in the HapMap2 CEU samples, 96.3% were observed in the Icelandic whole-genome sequencing data. Of the 6.9 million SNPs reported in the 1000 Genomes Project data, 89.4% were observed in the Icelandic whole-genome sequencing data.

Statistical analysis

Long range phasing. Long range phasing of all chip-genotyped individuals was performed with methods described previously[5-9]. In brief, phasing is achieved using an iterative algorithm which phases a single proband at a time given the available phasing information about everyone else that shares a long haplotype identically by state with the proband. Given the large fraction of the

Icelandic population that has been chip-typed, accurate long range phasing is available genome-wide for all chip-typed Icelanders.

Genotype imputation. We imputed the SNPs identified and genotyped through sequencing into all Icelanders who had been phased with long range phasing using the same model as used by IMPUTE[10]. The genotype data from sequencing can be ambiguous due to low sequencing coverage. In order to phase the sequencing genotypes, an iterative algorithm was applied for each SNP with alleles 0 and 1. We let H be the long range phased haplotypes of the sequenced individuals and applied the following algorithm:

1. For each haplotype h in H, use the Hidden Markov Model of IMPUTE to calculate for every other k in H, the likelihood, denoted $\gamma_{h,k}$, of h having the same ancestral source as k at the SNP. For every h in H, initialize the parameter θ_h, which specifies how likely the one allele of the SNP is to occur on the background of h from the genotype likelihoods obtained from sequencing. The genotype likelihood L_g is the probability of the observed sequencing data at the SNP for a given individual assuming g is the true genotype at the SNP. If L_0, L_1 and L_2 are the likelihoods of the genotypes 0, 1 and 2 in the individual that carries h, then set $\theta_h = \frac{L_2+\frac{1}{2}L_1}{L_2+L_1+L_0}$.

2. For every pair of haplotypes h and k in H that are carried by the same individual, use the other haplotypes in H to predict the genotype of the SNP on the backgrounds of h and k: $\tau_h = \sum_{l \in H \setminus \{h\}} \gamma_{h,l}\theta_l$ and $\tau_k = \sum_{l \in H \setminus \{k\}} \gamma_{k,l}\theta_l$. Combining these predictions with the genotype likelihoods from sequencing gives un-normalized updated phased genotype

probabilities: $P_{00} = (1 - \tau_h)(1 - \tau_k)L_0$, $P_{10} = \tau_h(1 - \tau_k)\frac{1}{2}L_1$, $P_{01} = (1 - \tau_h)\tau_k\frac{1}{2}L_1$ and $P_{11} = \tau_h\tau_k L_2$.

3. Now use these values to update θ_h and θ_k to $\theta_h = \frac{P_{10}+P_{11}}{P_{00}+P_{01}+P_{10}+P_{11}}$ and $\theta_k = \frac{P_{01}+P_{11}}{P_{00}+P_{01}+P_{10}+P_{11}}$.
4. Repeat step 3 when the maximum difference between iterations is greater than a convergence threshold ε. We used $\varepsilon=10^{-7}$.

Given the long range phased haplotypes and θ, the allele of the SNP on a new haplotype h not in H, is imputed as $\sum_{l \in H} \gamma_{h,l}\theta_l$.

The above algorithm can easily be extended to handle simple family structures such as parent-offspring pairs and triads by letting the P distribution run over all founder haplotypes in the family structure. The algorithm also extends trivially to the X-chromosome. If source genotype data are only ambiguous in phase, such as chip genotype data, then the algorithm is still applied, but all but one of the Ls will be 0. In some instances, the reference set was intentionally enriched for carriers of the minor allele of a rare SNP in order to improve imputation accuracy. In this case, expected allele counts will be biased toward the minor allele of the SNP. Call the enrichment of the minor allele E and let θ' be the expected minor allele count calculated from the naïve imputation method, and let θ be the unbiased expected allele count, then $\theta' = \frac{E\theta}{1-\theta+E\theta}$ and hence $\theta = \frac{\theta'}{E+(1-E)\theta'}$.

This adjustment was applied to all imputations based on enriched imputations sets. We note that if θ' is 0 or 1, then θ will also be 0 or 1, respectively.

In-silico genotyping. In addition to imputing sequence variants from the whole genome sequencing effort into chip genotyped individuals, we also performed a second imputation step where genotypes were imputed into relatives of chip genotyped individuals, creating *in-silico* genotypes. The inputs into the second imputation step are the fully phased (in particular every allele has been assigned a parent of origin) imputed and chip type genotypes of the available chip typed individuals. The algorithm used to perform the second imputation step consists of:

1. For each ungenotyped individual (the proband), find all chip genotyped individuals within two meiosis of the individual. The six possible types of two meiosis relatives of the proband are (ignoring more complicated relationships due to pedigree loops): Parents, full and half siblings, grandparents, children and grandchildren. If all pedigree paths from the proband to a genotyped relative go through other genotyped relatives, then that relative is excluded. E.g. if a parent of the proband is genotyped, then the proband's grandparents through that parent are excluded. If the number of meiosis in the pedigree around the proband exceeds a threshold (we used 12), then relatives are removed from the pedigree until the number of meiosis falls below 12, in order to reduce computational complexity.
2. At every point in the genome, calculate the probability for each genotyped relative sharing with the proband based on the autosomal SNPs used for phasing. A multipoint algorithm based on the hidden Markov model Lander-Green multipoint linkage algorithm using fast Fourier transforms is used to calculate these sharing probabilities[34,35]. First single point sharing probabilities are calculated by dividing the genome into 0.5cM bins and using the haplotypes over these bins as alleles. Haplotypes that are the same, except at most at a single SNP, are treated as identical. When the haplotypes in the pedigree are incompatible over a bin, then a uniform probability distribution was used for that bin. The

most common causes for such incompatibilities are recombinations in member belonging to the pedigree, phasing errors and genotyping errors. Note that since the input genotypes are fully phased, the single point information is substantially more informative than for unphased genotyped, in particular one haplotype of the parent of a genotyped child is always known. The single point distributions are then convolved using the multipoint algorithm to obtain multipoint sharing probabilities at the center of each bin. Genetic distances were obtained from the most recent version of the deCODE genetic map[6].

3. Based on the sharing probabilities at the center of each bin, all the SNPs from the whole genome sequencing are imputed into the proband. To impute the genotype of the paternal allele of a SNP located at x, flanked by bins with centers at x_{left} and x_{right}. Starting with the left bin, going through all possible sharing patterns v, let I_v be the set of haplotypes of genotyped individuals that share identically by descent within the pedigree with the proband's paternal haplotype given the sharing pattern v and $P(v)$ be the probability of v at the left bin – this is the output from step 2 above – and let e_i be the expected allele count of the SNP for haplotype i. Then $e_v = \frac{\sum_{i \in I_v} e_i}{\sum_{i \in I_v} 1}$ is the expected allele count of the paternal haplotype of the proband given v and an overall estimate of the allele count given the sharing distribution at the left bin is obtained from $e_{left} = \sum_v P(v) e_v$. If I_v is empty then no relative shares with the proband's paternal haplotype given v and thus there is no information about the allele count. We therefore store the probability that some genotyped relative shared the proband's paternal haplotype, $O_{left} = \sum_{v, I_v = \emptyset} P(V)$ and an expected allele count, conditional on the proband's paternal haplotype being shared by at least one genotyped relative: $c_{left} = \frac{\sum_{v, I_v \neq \emptyset} P(v) e_v}{\sum_{v, I_v \neq \emptyset} P(v)}$. In the same way calculate O_{right} and

c_{right}. Linear interpolation is then used to get an estimates at the SNP from the two flanking bins:

$$O = O_{left} + \frac{x - x_{left}}{x_{right} - x_{left}}\left(O_{right} - O_{left}\right),$$

$$c = c_{left} + \frac{x - x_{left}}{x_{right} - x_{left}}\left(c_{right} - c_{left}\right).$$

If θ is an estimate of the population frequency of the SNP then $Oc + (1 - O)\theta$ is an estimate of the allele count for the proband's paternal haplotype. Similarly, an expected allele count can be obtained for the proband's maternal haplotype.

Genotype imputation information. The informativeness of genotype imputation was estimated by the ratio of the variance of imputed expected allele counts and the variance of the actual allele counts:

$$\frac{Var(E(\theta|chip\ data))}{Var(\theta)},$$

where $\theta \in \{0, 1\}$ is the allele count. $Var(E(\theta|chip\ data))$ was estimated by the observed variance of the imputed expected counts and $Var(\theta)$ was estimated by $p(1 - p)$, where p is the allele frequency. For the present study, when imputed genotypes are used, the information value for all SNPs is between 0.92 and 0.99.

Case control association testing. Logistic regression was used to test for association between SNPs and disease, treating disease status as the response and expected genotype counts from imputation or allele counts from direct genotyping as covariates. Testing was performed using the likelihood ratio statistic. When testing for association based on the *in silico* genotypes, controls were matched to cases based on the informativeness of the imputed genotypes, such that for each case C controls of matching informativeness where chosen. Failing to match cases and controls will lead to a highly inflated genomic control factor, and in some cases may lead to spurious false positive findings. The informativeness of each of the imputation of each one of an individual's haplotypes was estimated by taking the average of

$$a(e,\theta) = \begin{cases} \frac{e-\theta}{1-\theta}, & e \geq \theta \\ \frac{\theta-e}{\theta}, & e < \theta \end{cases}$$

over all SNPs imputed for the individual, where e is the expected allele count for the haplotype at the SNP and θ is the population frequency of the SNP. Note that $a(\theta,\theta) = 0$ and $a(0,\theta) = a(1,\theta) = 1$. The mean informativeness values cluster into groups corresponding to the most common pedigree configurations used in the imputation, such as imputing from parent into child or from child into parent. Based on this clustering of imputation informativeness we divided the haplotypes of individuals into seven groups of varying informativeness, which created 27 groups of individuals of similar imputation informativeness; 7 groups of individuals with both haplotypes having similar informativeness, 21 groups of individuals with the two haplotypes having different informativeness, minus the one group of individuals with neither haplotype being imputed well. Within each group we calculate the ratio of the number of controls and the number of cases, and choose the largest integer C that was less than this ratio in all the groups. For example, if in one group there are 10.3 times as many controls as cases and if in all other groups this ratio was greater, then we would set $C = 10$ and within each group randomly select ten times as many controls as there are cases. For thyroid cancer we used $C = 109$ and for goiter we used $C = 186$.

For the 24 SNPs reported in Suppl. Table 2 we compared the correlation between the imputed and directly generated genotype. Based on a minimum of 1,545 individuals with both imputed and single track generated genotypes we calculated the correlation coefficient (r). The range of the correlation is between 0.92 and 1.00. The second lowest correlation value was 0.96 and the median value was 0.98.

Quantitative trait association testing. The normalized and adjusted measurements of TSH, FT_3 and FT_4were regressed on allele counts using classical linear regression.

Sibling recurrence risk ratio:

The sibling recurrence risk ratio is defined as $\lambda_{sibling} = \frac{P(A \mid B)}{P(A)} = \frac{P(AB)}{P(A)P(B)}$, where *A* is the event that a person gets a disease and *B* is the event that a particular sibling of the person gets the disease. Assuming a multiplicative model, the $\lambda_{\mathrm{sibling}}$ accounted for by a variant with frequency *f* and relative risk of *r* is equal to $\frac{\frac{1}{4}\left[fr^2 + 1 - f + (fr + 1 - f)^2\right]^2}{(fr + 1 - f)^4}$

Inflation factor adjustment. In order to account for the relatedness and stratification within our case and control sample sets we applied the method of genomic control based on chip markers. For the thyroid cancer GWAS the correction factor based on the genomic control is 1.14. For the GWAS on goiter, TSH, FT3 and FT4 the correction factor based on the genomic control is 1.08, 1.33, 1.02, and 1.16, respectively.

References

1. Kutyavin, I.V. et al. A novel endonuclease IV post-PCR genotyping system. *Nucleic Acids Research* **34**, e128 (2006).
2. He, H. et al. Allelic variation in gene expression in thyroid tissue. *Thyroid* **15**, 660-7 (2005).
3. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-60 (2009).
4. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-9 (2009).
5. Kong, A. et al. Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat Genet* **40**, 1068-75 (2008).
6. Kong, A. et al. Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* **467**, 1099-103 (2010).
7. Sulem, P. et al. Identification of low-frequency variants associated with gout and serum uric acid levels. *Nat Genet* **43**, 1127-30 (2011).
8. Rafnar, T. et al. Mutations in BRIP1 confer high risk of ovarian cancer. *Nat Genet* **43**, 1104-7 (2011).
9. Stacey, S.N. et al. A germline variant in the TP53 polyadenylation signal confers cancer susceptibility. *Nat Genet* **43**, 1098-103 (2011).
10. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* **39**, 906-13 (2007).