Global multi-ancestry genetic study elucidates genes and biological pathways associated with thyroid cancer and benign thyroid diseases

Samantha L. White

samantha.l.white@cuanschutz.edu

Department of Biomedical Informatics, University of Colorado Anschutz Medical Campus, Aurora, CO, USA

Maizy S. Brasher

maizy.brasher@cuanschutz.edu

Department of Biomedical Informatics, University of Colorado Anschutz Medical Campus, Aurora, CO, USA

Jack Pattee

pattee.jack@gmail.com

Center for Innovative Design & Analysis, Colorado School of Public Health,

University of Colorado Anschutz Medical Campus, Aurora, CO, USA

Wei Zhou

wzhou@broadinstitute.org

Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA, USA

Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, MA, USA

Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA

Sinéad Chapman

sinead@broadinstitute.org

The Broad Institute, Cambridge, MA, USA

Yon Ho Jee

jee@broadinstitute.org

Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA

Caitlin C. Bell

caitlin.c.bell@cuanschutz.edu

Division of Endocrinology, Diabetes and Metabolism, University of Colorado Anschutz Medical Campus, Aurora, CO, USA

Taylor L. Jamil

taylor.jamil@cuanschutz.edu

Department of Otolaryngology, Head and Neck Surgery, University of Colorado Anschutz Medical Campus, Aurora, Colorado, USA

Martin Barrio

martin.barrio@cuanschutz.edu

Division of GI, Trauma, and Endocrine Surgery, Department of Surgery, University of Colorado School of Medicine, Aurora, CO, USA

Jibril Hirbo

jibril.hirbo@vumc.org

Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA

Vanderbilt Genetic Institute, Vanderbilt University Medical Center, Nashville, TN, USA

Nancy J. Cox

nancy.j.cox@vumc.org

Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA

Vanderbilt Genetic Institute, Vanderbilt University Medical Center, Nashville, TN, USA

Peter Straub

peter.straub@gmail.com

Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USAz

Vanderbilt Genetic Institute, Vanderbilt University Medical Center, Nashville, TN, USA

Shinichi Namba

snamba@m.u-tokyo.ac.jp

Department of Genome Informatics, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan

Department of Statistical Genetics, Osaka University Graduate School of Medicine, Suita, Japan

Laboratory for Systems Genetics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan

Emily Bertucci-Richter

embricht@umich.edu

Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA

Lindsay Guare

lindsay.guare@pennmedicine.upenn.edu

Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

Ahmed EdrisMohammed

ahmed.edrismohamed@ndph.ox.ac.uk

Nuffield Department of Population Health, University of Oxford, Oxford, UK

Sam Morris

sam.morris@ndph.ox.ac.uk

Nuffield Department of Population Health, University of Oxford, Oxford, UK

Ashley J Mulford

amulford@northshore.org

Genomic Health Initiative, Endeavor Health Research Institute, Evanston, IL, USA

Haoyu Zhang

haoyu.zhang@cuanschutz.edu

Department of Biomedical Informatics, University of Colorado Anschutz Medical Campus, Aurora, CO, USA

Brian Fennessy

brian.fennessy@mssm.edu

The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA

Martin D Tobin

martin.tobin@leicester.ac.uk

Department of Population Health Sciences, University of Leicester, Leicester, UK

University Hospitals of Leicester NHS Trust, Leicester, LE15WW, UK

Jing Chen

jc824@leicester.ac.uk

Department of Population Health Sciences, University of Leicester, Leicester, UK

Alexander T Williams

atw20@leicester.ac.uk

Department of Population Health Sciences, University of Leicester, Leicester, UK

Catherine John

catherine.john@leicester.ac.uk

Department of Population Health Sciences, University of Leicester, Leicester, UK

University Hospitals of Leicester NHS Trust, Leicester, LE15WW, UK

David A van Heel

d.vanheel@qmul.ac.uk

Blizard Institute, Queen Mary University of London, London E1 2AB, UK

Rohini Mathur

r.mathur@qmul.ac.uk

Wolfson Institute of Population Health, Queen Mary University of London, London E1 2AB, UK

Sarah Finer

s.finer@qmul.ac.uk

Wolfson Institute of Population Health, Queen Mary University of London, London E1 2AB, UK

Marta Riise Moksnes

marta.r.moksnes@ntnu.no

HUNT Research Centre, Department of Public Health and Nursing, NTNU, Norwegian University of Science and Technology, Levanger, Norway

Ben Brumpton

ben.brumpton@ntnu.no

HUNT Research Centre, Department of Public Health and Nursing, NTNU, Norwegian University of Science and Technology, Levanger, Norway

Bjørn Olav Åsvold

bjorn.o.asvold@ntnu.no

HUNT Research Centre, Department of Public Health and Nursing, NTNU, Norwegian University of Science and Technology, Levanger, Norway

Raitis Peculis

raitis@biomed.lu.lv

Latvian Biomedical Research and Study Centre, Ratsupites 1-1, Riga, LV-1067, Latvia

Vita Rovite

vita.rovite@biomed.lu.lv

Latvian Biomedical Research and Study Centre, Ratsupites 1-1, Riga, LV-1067, Latvia

Ilze Konrade

ilze.konrade@rsu.lv

Department of Internal Medicine, Riga Stradins University, Dzirciema Str 16, LV-1007 Riga, Latvia

Ying Wang

yiwang@broadinstitute.org

Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA

Kristy Crooks

kristy.crooks@cuanschutz.edu

Colorado Center for Personalized Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO, USA

Sameer Chavan

sameer.chavan7@gmail.com

Colorado Center for Personalized Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO, USA

Matthew J. Fisher

matthew.fisher126@gmail.com

Colorado Center for Personalized Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO, USA

Nicholas Rafaels

nicholas.rafaels@cuanschutz.edu

Colorado Center for Personalized Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO, USA

Meng Lin

meng.lin@cuanschutz.edu

Department of Biomedical Informatics, University of Colorado Anschutz Medical Campus, Aurora, CO, USA

Jonathan Shortt

jonathan.shortt@cuanschutz.edu

Department of Biomedical Informatics, University of Colorado Anschutz Medical Campus, Aurora, CO, USA

Colorado Center for Personalized Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO, USA

Alan R Sanders

asanders@northshore.org

Genomic Health Initiative, Endeavor Health Research Institute, Evanston, IL, USA

Department of Psychiatry and Behavioral Neuroscience, University of Chicago, Chicago, IL, USA

David Whiteman

david.whiteman@qimrb.edu.au

QIMR Berghofer Medical Research Institute: Herston, QLD, AU

Stuart MacGregor

stuart.macgregor@qimrb.edu.au

QIMR Berghofer Medical Research Institute: Herston, QLD, AU

Sarah Medland

sarah.medland@qimrb.edu.au

QIMR Berghofer Medical Research Institute: Herston, QLD, AU

Unnur Thorsteinsdóttir

unnurth@decode.is

deCODE genetics/Amgen, Inc., Reykjavik, Iceland

Faculty of Medicine, University of Iceland, Reykjavik, Iceland

Kári Stefánsson

kstefans@decode.is

deCODE genetics/Amgen, Inc., Reykjavik, Iceland

Faculty of Medicine, University of Iceland, Reykjavik, Iceland

Tugce Karaderi

tugce.karaderi@sund.ku.dk

Center for Health Data Science, Section for Health Data Science and Artificial Intelligence, Department of Public Health, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark

Kathleen M. Egan

Kathleen.Egan@moffitt.org

Moffitt Cancer Center

Department of Cancer Epidemiology, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA

Therese Bocklage

Therese.Bocklage@uky.edu

University of Kentucky Markey Cancer Center

University of Kentucky-Department of Pathology and Laboratory Medicine, University of Kentucky-Chandler Medical Center, Lexington, KY, USA

Hilary C McCrary

Hilary.McCrary@hsc.utah.edu

University of Utah Huntsman Cancer Institute

Department of Otolaryngology-Head and Neck Surgery, School of Medicine, Huntsman Cancer Institute, University of Utah, Salt Lake City, UT, USA

Greg Riedlingeer

gr338@cinj.rutgers.edu

Rutgers Cancer Institute of New Jersey Department of Medicine, Division of Medical Oncology, Rutgers Cancer Institute of New Jersey, New Brunswick, NJ, USA

Bodour Salhia

salhia@usc.edu

Norris Comprehensive Cancer Center and Keck School of Medicine of University of Southern California

Department of Translational Genomics, Keck School of Medicine, University of Southern California, Los Angeles, California, USA

Craig Shriver

craig.shriver@usuhs.edu

Muthra Cancer Center

Murtha Cancer Center, Uniformed Services University/Walter Reed National Military Medical Center, Bethesda, MD, USA

Minh D Phan

minh-phan@ouhsc.edu

University of Oklahoma Stephenson Cancer Center

Medical Oncology/Head and Neck Oncology, Stephenson Cancer Center, University of Oklahoma, Oklahoma City, Oklahoma, USA

Janice L. Farlow

JLFarlow@iu.edu

Indiana University School of Medicine, Indianapolis, Indiana

Stephen Edge

Stephen.Edge@RoswellPark.org

Roswell Park Comprehensive Cancer Center

Departments of Surgical Oncology and Cancer Prevention and Control , Roswell Park Comprehensive Cancer Center , Buffalo , NY , USA

Varinder Kaur

VK4Q@uvahealth.org

University of Virginia Cancer Center

Department of Internal Medicine, Division of Hematology & Oncology, University of Virginia Health, Charlottesville, VA, USA

Michelle Churchman

Michelle.Churchman@asterinsights.com

Aster Insights, Hudson, FL, USA

Robert J. Rounbehler

Rob.Rounbehler@asterinsights.com

Aster Insights, Hudson, FL, USA

Pamela L. Brock

Pamela.Brock@osumc.edu

The Ohio State University Comprehensive Cancer Center and the Ohio State University College of Medicine, Columbus, OH, USA

Matthew D. Ringel

Matthew.Ringel@osumc.edu

The Ohio State University Comprehensive Cancer Center and the Ohio State University College of Medicine, Columbus, OH, USA

Milton Pividori

milton.pividori@cuanschutz.edu

Department of Biomedical Informatics, University of Colorado Anschutz Medical Campus, Aurora, CO, USA

Rebecca Schweppe

rebecca.schweppe@cuanschutz.edu

Division of Endocrinology, Diabetes and Metabolism, University of Colorado Anschutz Medical Campus, Aurora, CO, USA

University of Colorado Cancer Center, University of Colorado Anschutz Medical Campus, Aurora, CO, USA

Christopher D. Raeburn

christopher.raeburn@cuanschutz.edu

Division of GI, Trauma, and Endocrine Surgery, Department of Surgery, University of Colorado School of Medicine, Aurora, CO, USA

Robin Walters

robin.walters@ndph.ox.ac.uk

Nuffield Department of Population Health, University of Oxford, Oxford, UK

Zhengming Chen

Zhengming.chen@ndph.ox.ac.uk

Nuffield Department of Population Health, University of Oxford, Oxford, UK

Liming Li

lmlee@vip.163.com

Department of Epidemiology and Biostatistics, School of Public Health, Peking University, Beijing, China

Peking University Center for Public Health and Epidemic Preparedness and Response, Beijing, China

Key Laboratory of Epidemiology of Major Diseases (Peking University), Ministry of Education, Beijing 100191, China

Koichi Matsuda

kmatsuda@edu.k.u-tokyo.ac.jp

Laboratory of Genome Technology, Human Genome Center, Institute of Medical Science, The University of Tokyo, Tokyo, Japan

Laboratory of Clinical Genome Sequencing, Graduate School of Frontier Sciences, The University of Tokyo, Tokyo, Japan

Yukinori Okada

yuki-okada@m.u-tokyo.ac.jpDepartment of Genome Informatics, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan

Department of Statistical Genetics, Osaka University Graduate School of Medicine, Suita, Japan

Laboratory for Systems Genetics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan

Sebastian Zoellner

szoellne@umich.edu

Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA

Anurag Verma

anurag.verma@pennmedicine.upenn.edu

Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

Michael Preuss

michael.preuss@mssm.edu

The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA

Eimear Kenny

eimear.kenny@mssm.edu

The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA

Audrey Hendricks

audrey.hendricks@cuanschutz.edu

390 Department of Biomedical Informatics, University of Colorado Anschutz Medical
391 Campus, Aurora, CO, USA

392

393 **Lauren Fishbein**

394 lauren.fishbein@cuanschutz.edu

395 Division of Endocrinology, Diabetes and Metabolism, University of Colorado Anschutz
396 Medical Campus, Aurora, CO, USA

397 Department of Biomedical Informatics, University of Colorado Anschutz Medical
398 Campus, Aurora, CO, USA

399 University of Colorado Cancer Center, University of Colorado Anschutz Medical
400 Campus, Aurora, CO, USA

401 Research Service, Rocky Mountain Regional VA Medical Center, Aurora, CO, USA

402

403 **Peter Kraft**

404 phillip.kraft@nih.gov

405 Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA,
406 USA

407 Transdivisional Research Program, Division of Cancer Epidemiology and Genetics,
408 National Cancer Institute, National Institutes of Health, MD, USA

409

410 **Mark Daly**

411 mjdaly@broadinstitute.org

412 Program in Medical and Population Genetics, Broad Institute of MIT and Harvard,
413 Cambridge, MA, USA

414 Analytical and Translational Genetics Unit, Department of Medicine, Massachusetts
415 General Hospital and Harvard Medical School, Boston, MA, USA

416 Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard,
417 Cambridge, MA, USA

418

419 **Benjamin Neale**

420 bneale@broadinstitute.org

Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA

Analytical and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA

Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA

The biobank at the Colorado Center for Personalized Medicine

University of Colorado Anschutz Medical Campus, Aurora, CO, USA

Genes & Health Research Team

The BioBank Japan Project

Alicia Martin

armartin@broadinstitute.org

Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA , USA

Stanley Center for Psychiatric Research and Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA

Department of Medicine, Harvard Medical School, Boston, MA, USA

Joanne B. Cole

joanne.cole@cuanschutz.edu

Department of Biomedical Informatics, University of Colorado Anschutz Medical Campus, Aurora, CO, USA

Human Medical Genetics and Genomics Program, University of Colorado Anschutz Medical Campus, Aurora, CO, USA

Bryan R. Haugen

bryan.haugen@cuanschutz.edu

Division of Endocrinology, Diabetes and Metabolism, University of Colorado Anschutz Medical Campus, Aurora, CO, USA

University of Colorado Cancer Center, University of Colorado Anschutz Medical Campus, Aurora, CO, USA

Christopher R. Gignoux

chris.gignoux@cuanschutz.edu

Human Medical Genetics and Genomics Program, University of Colorado Anschutz Medical Campus, Aurora, CO, USA

Department of Biomedical Informatics, University of Colorado Anschutz Medical Campus, Aurora, CO, USA

Colorado Center for Personalized Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO, USA

University of Colorado Cancer Center, University of Colorado Anschutz Medical Campus, Aurora, CO, USA

Nikita Pozdeyev

nikita.pozdeyev@cuanschutz.edu

Department of Biomedical Informatics, University of Colorado Anschutz Medical Campus, Aurora, CO, USA

Division of Endocrinology, Diabetes and Metabolism, University of Colorado Anschutz Medical Campus, Aurora, CO, USA

University of Colorado Cancer Center, University of Colorado Anschutz Medical Campus, Aurora, CO, USA

Colorado Center for Personalized Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO, USA

Corresponding authors: Nikita Pozdeyev and Christopher Gignoux

E-mail: nikita.pozdeyev@cuanschutz.edu; chris.gignoux@cuanschutz.edu

Address: 1890 N Revere Ct. MS F600, Aurora, CO 80045

Phone: 303-724-9659

Nikita Pozdeyev's ORCiD ID: 0000-0001-8574-1972

Christopher Gignoux's ORCiD ID: 0000-0001-9728-6567

Abstract

Thyroid diseases are common and highly heritable. Under the Global Biobank Meta-analysis Initiative, we performed a meta-analysis of genome-wide association studies from 19 biobanks for five thyroid diseases: thyroid cancer, benign nodular goiter, Graves' disease, lymphocytic thyroiditis, and primary hypothyroidism. We analyzed genetic association data from ~2.9 million genomes and identified 235 known and 501 novel independent variants significantly linked to thyroid diseases. We discovered genetic correlations between thyroid cancer, benign nodular goiter, and autoimmune thyroid diseases (r^2=0.21-0.97). Telomere maintenance genes contribute to benign and malignant thyroid nodular disease risk, whereas cell cycle, DNA repair, and DNA damage response genes are predominantly associated with thyroid cancer. We proposed a paradigm explaining genetic predisposition to benign and malignant thyroid nodules. We evaluated thyroid cancer polygenic risk scores (PRS) for clinical applications in thyroid cancer diagnosis. We found PRS associations with thyroid cancer risk features: multifocality, lymph node metastases, and extranodal extension.

Thyroid diseases are highly prevalent. According to the American Thyroid Association, over 12 percent of the US population will develop a thyroid condition during their lifetime (https://www.thyroid.org/media-main/press-room/). Thyroid cancer is the most common endocrine malignancy, with 44,020 new cases and 2,170 deaths in the United States in 2024 [1]. Thyroid function diseases, hypo- and hyperthyroidism, negatively affect most organ systems and are associated with major excess in cardiovascular mortality [2]. It is not well understood why some individuals develop thyroid disease, although genetic [3,4] and environmental factors (radiation exposure[5]) play a role.

Genetic effects are estimated to contribute up to 53% to thyroid cancer susceptibility in family studies [3,4], making thyroid cancer one of the most heritable common cancers[3,6]. For autoimmune thyroid diseases, genetic factors account for approximately 75% of the total phenotypic variance [7].

Ruling out thyroid malignancy is a common clinical task due to the high prevalence of thyroid nodules. Thyroid ultrasound reveals nodules in up to 65% of the general population, with prevalence increasing in women, with increasing age, and after radiation exposure [8,9]. Clinical providers assess thyroid nodule sonographic characteristics [10] to decide if fine needle aspiration (FNA) biopsy is indicated. Over 600,000 fine needle aspiration thyroid nodule biopsies (FNA) are performed annually in the United States to rule out cancer [11], and most (~92%) produce benign, inadequate, or indeterminate results [12,13]. Genetic thyroid cancer risk assessment with polygenic risk score (PRS) provides an opportunity to improve the diagnostic yield of FNA and reduce unnecessary procedures, molecular tests and diagnostic surgeries [14].

In addition to the clinical applications, discovering genetic variants predisposing to thyroid cancer and benign thyroid conditions helps in understanding the biological processes leading to disease. Several genome-wide association studies (GWAS) have been conducted on thyroid cancer [15-20]. More recently, the Global Biobank Meta-analysis (GBMI) consortium combined data from 6,699 individuals with thyroid cancer and ~2.2 million controls and identified 27 significant genetic associations[21]. GWAS for benign thyroid diseases and related traits, such as thyroid-stimulating hormone (TSH) levels, have been performed in large biobanks: UK Biobank[22,23], FinnGen [24], and Million Veteran Program [25]. However, a systematic analysis of underlying genes, pathways, and clinical relevance is missing.

Platforms such as the Global Biobank Meta-analysis initiative (GBMI, https://www.globalbiobankmeta.org/ [21]) enable global collaborations among dozens of participating biobanks, resulting in unmatched GWAS discovery power and data diversity, particularly relevant to cross-phenotype investigations. In this study, we report results from a GBMI project dedicated to thyroid diseases. We confirmed many previously reported variants and discovered hundreds of novel genetic associations with thyroid diseases. We found genes and pathways unique to malignant and benign nodular disease. We defined potential clinical applications for polygenic risk scores

(PRS) for thyroid cancer to better distinguish thyroid cancer from benign nodules before FNA. We found a genetic predisposition to multifocal and metastatic thyroid cancer.

Results

The study had three phases (**Figure 1**): 1) Variant Discovery: GWAS, meta-analysis, and quality control procedures; 2) Functional Inference: genetic correlations, transcriptome-wide association studies (TWAS), pathway and gene expression analyses; and 3) Clinical Studies: PRS development, testing on the clinical use case of distinguishing benign from malignant thyroid nodules, and associations with cancer aggressiveness.

Virtual Thyroid Biopsy Consortium

We founded the Virtual Thyroid Biopsy Consortium (**Extended Data Figure 1**) under the Global Biobank Meta-analysis initiative (https://www.globalbiobankmeta.org/) [21] to study the genetic architecture of thyroid diseases at a global multi-ancestry scale. The Consortium aggregates data from 19 biobanks in 10 countries and four continents. (**Supplementary Table 1**). Biobanks performed multi-ancestry and/or ancestry-stratified GWAS for five thyroid diseases: thyroid cancer, benign nodular goiter, Graves' disease, lymphocytic thyroiditis and primary hypothyroidism. In addition, a GWAS of thyroid cancer vs. benign nodular goiter was performed, focusing on the common clinical task of determining malignancy in thyroid nodules. Phenotype and GWAS definitions are listed in **Supplementary Tables 2 and 3**.

Meta-analysis of genome-wide association studies for thyroid diseases

Our meta-analysis aggregated data from 198 GWAS summary data files (**Supplementary Table 4**). Individual GWAS runs were well controlled for confounding (covariate-adjusted LD score regression [cov-LDSC][26] Y-axis intercept 1.00 ± 0.05 [mean ± SD]). Healthcare system-based biobanks had a higher disease prevalence than population-based biobanks (**Extended Data Figure 2**), as reported previously [21]. Bio*Me,* the All of Us Research Program (AoU), and the Million Veteran Program Biobanks had the most diverse participant pools as measured by Summix2 [27].

The meta-analysis included 21,816 cases of thyroid cancer, 68,987 cases of benign nodular goiter, 18,719 cases of Graves' disease, 18,331 cases of lymphocytic thyroiditis, 257,365 cases of primary hypothyroidism, and ~ 2.9 million controls (**Supplementary Table 3**). Population structure was determined with Summix2 [27] via mixture modeling of study-based allele frequencies compared to the gnomAD reference panel [28]. Seventeen percent of genotypes were from individuals of African (AFR-like),

4.4 % from Admixed American (AMR-like), 8.1% from East Asian (EAS-like), and 70.5% from European (EUR-like) ancestry.

We found 736 independent (loci separated by at least 500 kb) variants significantly (p-value ≤ 5e-8) associated with thyroid diseases, including mixed-ancestry and ancestry-stratified genetic associations (**Table 1, Supplementary Tables 5.1-5.6**). Of these, 235 variants were reported to the NHGRI-EBI Catalog[29] for thyroid traits (as of April 2024), and 501 genetic associations were novel. Most lead associations are in the introns of protein-coding genes (n = 356), followed by intergenic variants (n = 215 variants). Among 46 significant exonic variants, 43 are non-synonymous, potentially altering protein function.

Ancestry-stratified GWAS (**Supplementary Tables 5.1-5.6**) replicated many associations from the combined mixed-ancestry meta-analysis with some exceptions. For example, a rare (minor allele frequency [MAF] = 0.0007) non-synonymous exonic variant in the shelterin complex gene *TERF1* (8:73046129:G:A, β=1.33, p-value = 8.8e-10) was significantly associated with thyroid cancer in the EUR-like meta-analysis only (mixed ancestry GWAS β=1.16, p-value = 5e-4).

SNP heritability (h^2_{SNP}) and genetic correlation.

Cov-LDSC estimated h^2_{SNP} ranged from 0.067 (SE = 0.01) for benign nodular goiter in mixed-ancestry meta-analysis to 0.11 (0.026) for EUR-like thyroid cancer meta-analysis (**Supplementary Table 6**).

Genetic correlation was very high for lymphocytic thyroiditis and hypothyroidism (mixed-ancestry, r^2 = 0.97 [0.04], p-value = 7.9e-105, **Figure 2, Supplementary Table 7**). We found significant (Benjamini-Hochberg false discovery rate (FDR) <0.05) genetic correlations between lymphocytic thyroiditis and Graves' disease (r^2 = 0.62 [0.07]), lymphocytic thyroiditis and benign nodular goiter (r^2 = 0.15 [0.07]), thyroid cancer and benign nodular goiter (r^2 = 0.40 [0.16]), Graves' disease and hypothyroidism (r^2 = 0.36 [0.07]), Graves' disease and benign nodular goiter (r^2 = 0.31 [0.07]), and Graves' disease and thyroid cancer (r^2 = 0.21 [0.05]). Genetic correlation analysis in EUR-like meta-analysis showed similar results. (**Extended Data Figure 3, Supplementary Table 7**).

Transcriptome-wide association study (TWAS).

We performed *cis*-acting expression quantitative trait loci (*cis*-eQTLs) TWAS using two methods, FUSION[30] and S-PrediXcan[31,32] and GTEx v8 thyroid tissue expression models [33], to identify potential causal variants affecting gene expression and assign intergenic and non-coding RNA variants to protein-coding genes.

FUSION TWAS, as applied to the mixed-ancestry and EUR-like thyroid cancer GWAS meta-analysis, identified expression of 55 unique protein-coding genes (**Supplementary Tables 8 and 9**). FUSION also identified 47 and 45 significant (after

Bonferroni adjustment) lead *cis*-eQTL variants from mixed ancestry and EUR-like GWAS, respectively. TWAS attributed many significant intergenic and non-coding variants to protein-coding genes based on reported eQTL status. For example, GWAS intergenic variant 1:218515813:T:C (mixed-ancestry thyroid cancer GWAS meta-analysis p-value 4.07e-39) was also a lead *cis*-eQTL variant in FUSION TWAS for *TGFB2* (p-value = 3.59e-61). Most significant genes found by FUSION TWAS were also replicated by S-PrediXcan, indicating the analytic rigor of our analyses (**Supplementary Tables 5.1-5.6**).

TWAS found additional significant genes where meta-analysis failed to identify genome-wide significant associations, e.g., *VEGFC* (p-value = 1.30e-06), *CDCA7L* (p-value = 6.02e-07), and *NBR1* (p-value = 1.02e-06), further expanding our knowledge of genes associated with thyroid cancer risk beyond those with variants with marginal significance.

Gene expression analysis.

We evaluated mRNA expression of genes discovered in the thyroid cancer GWAS meta-analysis and TWAS in normal and malignant thyroid tissues (**Extended Data Figure 4**, **Supplementary Table 10**). Of the twenty total evaluated tissue types [34], normal thyroid tissue was among the top three highest-expressing tissues for twenty-two genes. Two genes, *TG* and *PTCSC2*, are expressed only in the thyroid.

The expression of four genes (*VAV3*, *PCNX2*, *ETS1*, and *PIBF1*) correlated with younger age at thyroid cancer diagnosis in The Cancer Genome Atlas study for papillary thyroid cancer (THCA-TCGA)[35] and/or the Oncology Research Information Exchange Network (ORIEN) AVATAR study (https://www.oriencancer.org/) (r^2 = -0.14 to -0.22, p-value ≤ 4.2e-05). *TERT* expression correlated with older age at diagnosis (r^2 = -0.17 to -0.21, p-value ≤ 7.4e-05), matching a similar association with somatic *TERT* promoter mutations [36]. Overall, the expression of most significant genes (47 out of 54 with RNA sequencing data in THCA-TCGA) was correlated with at least one clinical or molecular thyroid cancer risk feature: higher stage, presence of extrathyroidal extension, *BRAF* V600E mutation, lower BRAF/RAS score (indicating BRAF-like expression profile[35]), higher ERK score (measuring RAS/MAPK pathway activity), and lower thyroid cancer differentiation (estimated with thyroid differentiation score [35]) (**Extended Data Figure 4**, **Supplementary Table 10,** Bonferroni-corrected p-value ≤ 9.1e-05).

Pleiotropic and disease-specific associations with thyroid cancer and benign nodular goiter.

We do not know why some patients develop benign thyroid nodules, while others get thyroid cancer. To understand cellular functions and pathways leading to benign or malignant thyroid nodular disease, we explored pleiotropic and thyroid cancer- and benign nodular goiter-specific loci (**Figure 2**).

We generated locus plots for genes with significant GWAS meta-analysis associations and genes with significant *cis*-eQTL in TWAS (**Supplementary Figures 1.1-1.3**). We categorized loci as those significantly associated with: 1) thyroid cancer but not benign nodular goiter (these loci may contribute to malignant transformation of follicular cells, **Supplementary Table 11, Supplementary Figure 1.1**); 2) benign nodular goiter but not thyroid cancer (these loci may lead to non-neoplastic thyroid nodules and thyroid neoplasms with low malignant potential, **Supplementary Figure 1.2**); and 3) both benign and malignant thyroid nodules (**Supplementary Figure 1.3**). In addition, we identified loci that were strongly associated with thyroid cancer (p-value < 1e-20) but relatively weakly (p-value >1e-10) associated with benign nodular goiter (e.g., *HAUS6* and *SDCCAG*) despite greater discovery power of benign nodular goiter GWAS meta-analysis.

Among 28 genes associated predominantly with thyroid cancer, 12 encode components of cell cycle checkpoints and proteins regulating centrosome and kinetochore function, microtubule attachment, and chromosome segregation (*ATM, CDC25B, CENPB, CENPE, CEP120, CHEK2, HAUS6, HMMR, MAD1L1, NUF1, PMF1, SDCCAG8*). Five genes play a role in DNA repair and cellular response to DNA damage (*ATM, DCLRE1B, PCNX2, PML, TP53*).

Benign nodular goiter-specific genes (n = 38) participate in insulin-like growth factor 1 (*IGF1, IGF2BP2, and IRS1*) and fibroblast growth factor (*FGF7, FRS2*) signaling pathways. Genes participating in thyroid gland development and thyroid hormone synthesis were linked to benign nodules (*GLIS3, TPO*), but some are also associated with thyroid cancer (*NKX2-1, TG*).

Notably, significant loci in telomere maintenance genes (*ACTRT3, TERC, LRRC34, ETS1, EXO1, STN1, TERT*) were associated with both thyroid cancer and benign nodular goiter. Genes participating in apoptosis, senescence, and transforming growth factor β (TGFβ) signaling are present in all three gene categories (**Supplementary Table 11**) and contribute to the development of both benign and malignant thyroid nodules. Variants in some of these overlapping genes (e.g., *TERT,* 5:1280013:T:C, β = 0.11 [0.02], p-value = 1.71e-10; NRG1, 8:32571946:G:A, β = -0.18 [0.02], p-value = 9.89e-25) were also significant in our meta-analysis of thyroid cancer vs. benign nodular goiter GWAS (**Supplementary Table 5.6**), indicating differential contribution to these diseases. Of particular interest is the *PTCSC2* locus because its significant variants have the opposite direction of effect with thyroid cancer and benign nodular goiter (**Extended Data Figure 5**, ρ = -0.77, p-value = 1.2e-24).

KEGG[37] and Reactome[38] pathway analysis identified cell cycle, senescence and apoptosis as key biological processes contributing to thyroid cancer risk (**Supplementary Tables 12 and 13**). IGF1 and PI3K/AKT signaling pathways were significantly associated with benign nodular goiter. Consistent with our analysis of individual genes (**Supplementary Table 11, Supplementary Figure 1.3**) telomere-related pathways were found in both thyroid cancer and benign nodular goiter analyses.

706

707 **Pleiotropic and disease-specific associations with autoimmune thyroid**
708 **diseases.**

709 Graves' disease and lymphocytic thyroiditis/primary hypothyroidism are related
710 autoimmune endocrine diseases with opposite clinical manifestations, causing
711 hyperthyroidism and hypothyroidism, respectively [39].

712 Plausibly, most genes and KEGG and Reactome pathways associated with
713 autoimmune thyroid diseases (**Supplementary Tables 12 and 13**) are related to the
714 immune system. Almost all loci significantly associated with Graves' disease
715 (**Supplementary Figure 2.1**) are also linked to primary hypothyroidism. Four genes
716 (*CD2, CD40, LINGO2, and TNRC18*) were discovered in Graves' disease (p-value < 5e-
717 8) but not hypothyroidism GWAS meta-analysis (**Supplementary Figure 2.2**). Genetic
718 associations with lymphocytic thyroiditis (**Supplementary Table 5.4**) replicated those
719 with primary hypothyroidism (**Figure 2**).

720

721 **Polygenic risk score (PRS) for thyroid cancer diagnosis.**

722 PRS quantifies an individual's risk for developing a specific trait or disease based
723 on their genetics. We explored the ability of PRS to identify people at high risk for
724 thyroid cancer ($PRS_{ThC\ vs.\ All}$) in the Colorado Center for Personalized Medicine (CCPM)
725 Biobank population (n = 94,651). $PRS_{ThC\ vs.\ All}$ was calculated from the independent
726 significant variants from the multi-ancestry thyroid cancer meta-analysis, leaving CCPM
727 out of training to avoid overfitting.

728 Population screening for thyroid cancer is controversial; therefore, we also
729 assessed the utility of PRS for the clinically relevant task of distinguishing benign from
730 malignant thyroid nodules ($PRS_{ThC\ vs.\ BNG}$). $PRS_{ThC\ vs.\ BNG}$ was defined as the difference
731 between $PRS_{ThC\ vs.\ All}$ and PRS for benign nodular goiter ($PRS_{BNG\ vs.\ All}$): $PRS_{ThC\ vs.\ BNG} =$
732 $PRS_{ThC\ vs.\ All} - PRS_{BNG\ vs.\ All}$.

733 $PRS_{ThC\ vs.\ All}$ achieved an area under the receiver operating characteristic curve
734 (AUC) of 0.686 (95% CI, 0.669,0.703, **Supplementary Table 14**). Genome-wide
735 $PRS_{ThC\ vs.\ All}$ calculated using PRS-CS[40] showed an AUC of 0.695 (0.669,0.721, **Figure**
736 **3A**). Individuals with $PRS_{ThC\ vs.\ All}$ in the top decile have ~10.7 times the odds to develop
737 thyroid cancer than those in the first decile (**Figure 3B**).

738 Our $PRS_{ThC\ vs.\ All}$ significantly outperformed the thyroid cancer PRS derived from
739 the largest previous GWAS meta-analysis from the GBMI phase I project[21] (AUC 0.651
740 [0.632, 0.671], DeLong test p-value = 1.13e-07). This improvement highlights the
741 greater discovery power of a large meta-analysis.

742 To test PRS performance on a clinically relevant use case of discriminating
743 benign and malignant thyroid nodules (ThC vs. BNG), three clinicians (CB, TJ and NP)

744 performed clinical chart reviews. We confirmed the diagnosis of non-medullary thyroid
745 cancer in 1,344 patients and the diagnosis of benign nodular goiter in 281 (out-of-
746 sample test set). All benign cases were supported by surgical histopathology to avoid
747 contamination due to small thyroid cancers not eligible for biopsy.

748 $PRS_{ThC\ vs.\ All}$ performed worse for the clinical ThC vs. BNG task (AUC 0.618
749 (0.561,0.675), which is expected due to the genetic associations shared between
750 thyroid cancer and benign nodular goiter. $PRS_{ThC\ vs.\ BNG}$, leveraging genetic associations
751 with both thyroid cancer and benign nodular goiter, demonstrated an improved AUC for
752 the ThC vs. BNG clinical task (0.664 [0.586,0.742], DeLong test p-value = 5e-4).
753 Thyroid nodules in individuals with $PRS_{ThC\ vs.\ BNG}$ in the top decile had ~7.8 times the
754 odds to be malignant than in individuals with $PRS_{ThC\ vs.\ BNG}$ in the first decile (**Figure
755 3B**).

756 For benign thyroid diseases, PRS AUCs ranged from 0.585 (0.572,0.599) for
757 benign nodular goiter to 0.664 (0.638,0.69) for Graves' disease. PRS analyses in the
758 European population showed results similar to those from the mixed-ancestry GWAS
759 meta-analysis (**Supplementary Table 14**).

760 Incorporating demographic and genetic ancestry covariates improved predictions
761 for thyroid cancer ($PRS_{ThC\ vs.\ All}$ AUC 0.721 [0.707,0.735]) and other thyroid diseases
762 (AUC ranging from 0.686 (0.655,0.718) for benign nodular goiter to 0.727 (0.709,0.745)
763 for hypothyroidism). We expected this improvement because of the higher incidence of
764 thyroid diseases in women[41] and the increased risk of developing thyroid nodules and
765 hypothyroidism with age[8,42]. However, no significant improvement in clinical $PRS_{ThC\ vs.}$
766 $_{BNG}$ performance was observed (**Supplementary Table 14**).

767 We did not find a significant drop in PRS performance measured with AUC in
768 EUR-like, AMR-like, and AFR-like strata (De-Long test, p-value > 0.05) except for
769 hypothyroidism PRS in AFR-like individuals (**Extended Data Figure 6**).

770

771 **Polygenic risk score and thyroid cancer aggressiveness.**

772 We evaluated associations between thyroid cancer PRS and aggressive features
773 of thyroid cancer in three domains (patient, tumor, and metastatic disease), abstracted
774 from surgical histopathology reports and clinical notes (**Figure 3C, Supplementary
775 Table 15**). $PRS_{ThC\ vs.\ All}$ was significantly associated with tumor focality and extranodal
776 extension. $PRS_{ThC\ vs.\ BNG}$ was significantly associated with tumor focality and the number
777 of neck lymph node metastases (p-value ≤ 1.7e-3, Bonferroni-adjusted p-value
778 threshold). In addition, $PRS_{ThC\ vs.\ All}$ was associated with lymphatic invasion, and PRS_{ThC}
779 $_{vs.\ BNG}$ was related to the risk of structural disease recurrence (defined per the American
780 Thyroid Association guidelines [43]) at a nominal p-value threshold of ≤0.05.

781

Discussion

We completed the largest GWAS meta-analysis for five thyroid diseases, leveraging a global collaboration involving 19 Biobanks from 10 countries. The Consortium replicated 235 genetic associations deposited in the NHGRI-EBI GWAS Catalog as of April 2024 (v.1.0.2) and discovered 501 new independent associations (**Table 1**).

Genetic correlation analysis (**Figure 2**) identified shared genetic architecture between thyroid diseases that is physiologically plausible and clinically meaningful. Chronic lymphocytic thyroiditis is a leading cause of primary hypothyroidism[44], explaining the near-perfect genetic correlation between these two diseases. The shared genetic basis for lymphocytic thyroiditis and Graves' disease is also expected because both conditions are autoimmune diseases with highly concordant familial risk [45]. The genetic correlation between Graves' disease and thyroid nodular disease (both benign nodules and thyroid cancer) may be mechanistically explained by enhanced thyroid-stimulating hormone (TSH) receptor signaling, which promotes thyroid epithelial growth and protects thyroid cells from apoptosis [46]. Previous population-based studies found an increased risk of thyroid (hazard ratio 10-15) and other cancers in patients with Graves' disease [47], consistent with our findings.

Shared (genetic correlation r^2 = 0.4-0.5) and unique genetic associations with thyroid cancer and benign nodular goiter allowed insights into genes and pathways that lead to malignant and benign thyroid nodules. Our hypothesis explaining why some individuals are susceptible to thyroid nodules while others develop thyroid cancer is shown in **Figure 4**. We propose that two biological processes with distinct genetic architecture cause thyroid nodules: 1) hyperplasia, a polyclonal follicular cell proliferation with no malignant potential, and 2) neoplasia, a clonal growth driven by somatic genetic alterations. Neoplastic nodules can be benign or malignant, resulting in the mismatch between biological mechanisms (hyperplasia vs. neoplasia) and GWAS phenotype definitions (benign and malignant thyroid nodules), resulting in partial overlap in genetic associations and significant genetic correlation between thyroid cancer and benign nodular goiter phenotypes.

We found that genes participating in the cell cycle, DNA repair and cellular response to DNA damage are predominantly associated with thyroid cancer but not benign nodules, highlighting the importance of these biological processes for malignant transformation of thyroid follicular cells. These variants and genes can lead to more aggressive multifocal and metastatic thyroid cancer (**Figure 3C**). On the contrary, genes in fibroblast growth factor and insulin-like growth factor 1 signaling pathways were uniquely associated with benign nodular goiter and may lead to hyperplastic benign thyroid nodules without malignant potential. Variants in genes participating in telomere maintenance increase the risk of thyroid cancer and benign neoplastic thyroid nodules (adenomas). It was demonstrated previously that telomere-lengthening germline variants predispose to papillary thyroid cancer. [48]

Our finding that autoimmune thyroid disorders share most genetic associations (**Figure 2**) indicates that similar fundamental mechanisms lead to Graves' disease and lymphocytic thyroiditis/primary hypothyroidism despite opposite clinical manifestations.

Of special interest are genes that were only found in Graves' disease meta-analysis despite the much greater discovery power of hypothyroidism GWAS. These unique associations (*CD2, CD40, etc.*) may be involved in the immune system processes that define the type of autoantibodies produced: TSHR antibodies in Graves' disease or TPO/TG antibodies in lymphocytic thyroiditis and primary hypothyroidism. Consistently, variants in *TSHR* was strongly associated with Graves' disease (e.g., 14:80990913:A:C, β=0.27 [0.01], p-value = 2.52e-137), while *TPO* and *TG* associations were only seen in hypothyroidism meta-analysis.

Thyroid cancer caused 2,170 deaths in the United States in 2024[1]. PRS derived from thyroid cancer GWAS can identify individuals at thyroid cancer risk in the population (**Figure 4A** and **B**, and [49]). Thyroid cancer screening is not currently recommended by the US Preventive Services Task Force (USPSTF) [50] due to concerns about overtreatment and lack of mortality benefit. However, we found that PRS is associated with high-risk thyroid cancer features (**Figure 4C**). Genetically-informed screening, when thyroid ultrasound is recommended only for a subset of individuals with high PRS, will not only increase diagnostic yield but may help identify patients who will benefit from early diagnosis and screening. Clinical testing of such an approach is needed.

Another clinically meaningful application for thyroid cancer PRS is for aiding in the diagnosis of thyroid cancer in patients with thyroid nodules [14]. Despite the widespread use of clinical ultrasound-based algorithms [8,9], 72% of FNAs produce benign results, 20% are inadequate or indeterminate [12,13,51,52]. PRS provides cancer risk assessment that is complementary and synergistic to ultrasound-based nodule evaluation [14].

We found that PRS derived from thyroid cancer GWAS meta-analysis alone is inferior for distinguishing benign and malignant thyroid nodules (because many genetic associations are shared between these two phenotypes, **Supplementary Table 14**). Incorporating variants from both thyroid cancer and benign nodular goiter meta-analysis ($PRS_{ThC\ vs.\ BNG}$) improved PRS performance. Active surveillance of thyroid nodules with low-risk sonographic appearance in patients with reassuring PRS could reduce the need for invasive procedures. Other putative clinical applications for PRS that have not yet been explored are for molecular diagnostics of thyroid nodules with indeterminate FNA cytology and to personalize decisions to proceed with hemi- or total thyroidectomy for thyroid cancer.

PRSs for autoimmune thyroid diseases have significant predictive power (e.g., covariate-adjusted hypothyroidism PRS AUC of 0.721 [0.701, 0.742]). However, both

hypothyroidism and Graves' disease are reliably diagnosed with biochemical and antibody testing, and the utility of genetic risk evaluation is uncertain.

We recognize that, due to the demographics of participants in the Virtual Thyroid Biopsy Consortium, we are underpowered in our ability to study individuals of non-EUR-like ancestry. As our consortium grows, we look forward to conducting more ancestry-specific analyses to ensure we find the results relevant to all individuals[53], and improve our insights into rare variation across groups.

In summary, we conducted the largest meta-analysis of GWAS for five thyroid diseases. We found many previously known and novel mechanistically plausible variants, genes, and pathways contributing to the risk of thyroid cancer, benign nodular goiter, and autoimmune thyroid diseases. We explained why some individuals are prone to developing benign thyroid nodules while others are at risk of multifocal metastatic thyroid cancer. We derived and tested polygenic risk scores for thyroid cancer population screening and for a more clinically relevant task of distinguishing benign and malignant thyroid nodules. This study will serve as a foundation for future clinical applications leveraging the germline genetics of thyroid diseases.

Acknowledgments.

We want to acknowledge the participants and investigators of all participating biobanks. We thank Regeneron Genetics Center for genotyping part of the Colorado Center for Personalised Medicine data.

We gratefully acknowledge All of Us participants for their contributions, without whom this research would not have been possible. We also thank the National Institutes of Health's All of Us Research Program for making available the participant data examined in this study.

The authors would like to acknowledge the Oncology Research Information Exchange Network (ORIEN) Institution Members for their commitment to data sharing. Thyroid cancer gene expression data included in this work was obtained through the Avatar Project.

Genes & Health thanks Social Action for Health, Centre of The Cell, members of our Community Advisory Group, and staff who have recruited and collected data from volunteers. We thank the NIHR National Biosample Centre (UK Biocentre), the Social Genetic & Developmental Psychiatry Centre (King's College London), Wellcome Sanger Institute, and Broad Institute for sample processing, genotyping, sequencing and variant annotation. This work uses data provided by patients and collected by the NHS as part of their care and support. This research utilised Queen Mary University of London's

Apocrita HPC facility, supported by QMUL Research-IT, http://doi.org/10.5281/zenodo.438045

Genes & Health thanks Barts Health NHS Trust, NHS Clinical Commissioning Groups (City and Hackney, Waltham Forest, Tower Hamlets, Newham, Redbridge, Havering, Barking and Dagenham), East London NHS Foundation Trust, Bradford Teaching Hospitals NHS Foundation Trust, Public Health England (especially David Wyllie), Discovery Data Service/Endeavour Health Charitable Trust (especially David Stables), Voror Health Technologies Ltd (especially Sophie Don), NHS England (for what was NHS Digital) - for GDPR-compliant data sharing backed by individual written informed consent.

Funding.

This project was funded by the National Cancer Institute grant 1R21CA282380 to Nikita Pozdeyev, Bryan Haugen and Christopher Gignoux, and the Colorado Clinical and Translational Sciences Institute grant CO-J-24-170 to Nikita Pozdeyev.

Lauren Fishbein is supported by the VA Merit Award I01BX006252.

Tugce Karaderi is supported by the Novo Nordisk Foundation Data Science Emerging Investigator grant (NNF20OC0062294).

Milton Pividori is supported by the National Human Genome Research Institute (R00HG011898).

Genes & Health is/has recently been core-funded by Wellcome (WT102627, WT210561), the Medical Research Council (UK) (M009017, MR/X009777/1, MR/X009920/1), Higher Education Funding Council for England Catalyst, Barts Charity (845/1796), Health Data Research UK (for London substantive site), and research delivery support from the NHS National Institute for Health Research Clinical Research Network (North Thames). We acknowledge the support of the National Institute for Health and Care Research Barts Biomedical Research Centre (NIHR203330); a delivery partnership of Barts Health NHS Trust, Queen Mary University of London, St George's University Hospitals NHS Foundation Trust and St George's University of London. Genes & Health is/has recently been funded by Alnylam Pharmaceuticals, Genomics PLC; and a Life Sciences Industry Consortium of AstraZeneca PLC, Bristol-Myers Squibb Company, GlaxoSmithKline Research and Development Limited, Maze Therapeutics Inc, Merck Sharp & Dohme LLC, Novo Nordisk A/S, Pfizer Inc, Takeda Development Centre Americas Inc.

Resource availability.

The meta-analysis GWAS summary statistics will be deposited in the GWAS Catalog after the acceptance of this manuscript for peer-reviewed publication. The PRS weights will be deposited in the PGS Catalog (https://www.pgscatalog.org/). All original code is publicly available at GitHub at https://github.com/pozdeyevlab/gwas-analysis. Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Nikita Pozdeyev, email: nikita.pozdeyev@cuanschutz.edu.

Disclosures.

Nikita Pozdeyev and Bryan R. Haugen receive research support from Veracyte, Inc., unrelated to this study.

Contributions.

Nikita Pozdeyev and Christopher R. Gignoux conceived, designed and supervised the project. Samantha L. White and Nikita Pozdeyev conducted the main analyses in this study. Nikita Pozdeyev and Samantha L. White wrote the initial draft of the paper. Maizy S. Brasher and Joanne B. Cole contributed data and analyses from the UK Biobank. Jibril Hirbo, Nancy J. Cox, and Peter Straub contributed data and analyses from the BioVU. Shinichi Namba, Koichi Matsuda and Yukinori Okada contributed data and analyses from the BioBank Japan Project. Emily Bertucci-Richter and Sebastian Zoellner contributed data and analyses from the Michigan Genomics Initiative. Lindsay Guare and Anurag Verma contributed data and analyses from the Penn Medicine BioBank. Ahmed EdrisMohammed, Sam Morris, Robin Walters and Zhengming Chen contributed data and analyses from the China Kadoorie Biobank. Ashley J Mulford and Alan R Sanders contributed data and analyses from the Genomics Health Initiative. Brian Fennessy, Michael Preuss and Eimear Kenny contributed data and analyses from the BioMe. Jack Pattee, Sameer Chavan, Matthew J. Fisher, Nicholas Rafaels, Meng Lin, Jonathan Shortt and Kristy Crooks contributed data and analyses from the biobank at the Colorado Center for Personalized Medicine. Yon Ho Jee, Peter Kraft and Alicia Martin contributed data and analyses from the Korean Cancer Prevention Study-II. Martin D Tobin, Jing Chen, Alexander T, Williams and Catherine John contributed data and analyses from the EXCEED study. David A. van Heel, Rohini Mathur and Sarah Finer contributed data and analyses from the Genes and Health study. Marta Riise Moksnes, Ben Brumpton, and Bjørn Olav Åsvold contributed data and analyses from the Trøndelag Health Study. Raitis Peculis, Vita Rovite, and Ilze Konrade contributed data and analyses from the Latvian Biobank. Ying Wang contributed data and analyses from the Mass General Brigham Biobank. David Whiteman, Stuart MacGregor and Sarah Medland contributed data and analyses from the QSkin Sun and Health Study. Unnur Thorsteinsdóttir and Kári Stefánsson contributed data and analyses from

deCODE Genetics. Lauren Fishbein contributed data and analyses from the Million Veteran Program. Kathleen M. Egan, Therese Bocklage, Hilary C McCrary, Greg Riedlingeer, Bodour Salhia, Craig Shriver, Minh D Phan, Janice L. Farlow, Stephen Edge, Varinder Kaur, Michelle Churchman, Robert J. Rounbehler, Pamela L. Brock, and Matthew D. Ringel contributed the Oncology Research Information Exchange Network (ORIEN) AVATAR study data. Haoyu Zhang and Milton Pividori performed S-PrediXcan analyses. Caitlin C. Bell, Taylor L. Jamil, Martin Barrio, Christopher D. Raeburn and Bryan R. Haugen collected and interpreted thyroid cancer clinical data. Tugce Karaderi contributed to the interpretation of the hypothyroidism meta-analysis data. Rebecca Schweppe assisted with the interpretation of pathway analysis data. Audrey Hendricks advised on the use of Summix2 for population structure estimates. Wei Zhou, Sinéad Chapman, Mark Daly and Benjamin Neale led the Global Biobank Meta-analysis Initiative. All authors contributed and approved the final version of the paper.

Methods

Ethical approval

Colorado Multiple Institutional Review Board for the University of Colorado Denver Anschutz Medical Campus, Aurora, Colorado, USA, waived ethical approval for this work (COMIRB #20-2315).

This research has been conducted using the UK Biobank Resource under Application Number 95339.

Virtual Thyroid Biopsy Consortium

We founded the Virtual Thyroid Biopsy (VTB) Consortium under the umbrella of the Global Biobank Meta-analysis Initiative (GBMI) [21]. Nineteen biobanks from ten countries and four continents contributed GWAS results to the meta-analysis (**Extended Data Figure 1**). **Supplementary Table 1** lists biobank size, available ancestry strata, phenotyping, genotyping and imputation methods, and software for GWAS.

Phenotype definitions

We defined thyroid phenotypes using International Classification of Diseases ICD-9-CM and ICD-10-CM billing codes for United States biobanks, ICD-9 and ICD-10 billing codes for international biobanks, and SNOMED codes and survey codes for the All of Us Research Program biobank (**Supplementary Table 2**). These phenotype definitions were shared with teams participating in the Consortium.

To evaluate polygenic risk score performance and study its association with thyroid cancer risk phenotypes, we performed clinical chart reviews for genotyped participants in the Colorado Center of Personalized Medicine (CCPM) Biobank.

1011 Histopathologic and cytologic diagnosis, patient characteristics (age at thyroid cancer
1012 diagnosis, death from thyroid cancer and risk of structural disease recurrence), tumor
1013 characteristics (tumor size, tumor focality, presence of extrathyroidal extension,
1014 lymphatic and angioinvasion, surgical margins positivity) and metastatic disease
1015 characteristics (presence of locoregional and distant metastases, extranodal extension,
1016 size and number of lymph node metastases) were extracted from surgical
1017 histopathology reports, thyroid nodule fine-needle aspirations reports, and
1018 endocrinology notes.

1019 The risk of structural disease recurrence was estimated on a continuous scale (1-
1020 55% risk) as described in the American Thyroid Association thyroid cancer guidelines.[43]
1021 For patients with multiple surgeries, the highest stage/risk was used (e.g., if the first
1022 surgery histopathology evaluation reported Nx stage but lateral neck metastases were
1023 found later, N1b stage was used for the association analysis). Thyroid cancer risk
1024 phenotypes are summarized in **Supplementary Table 14**. Benign cases for PRS
1025 evaluation in CCPM were defined based on surgical histopathology reports.

1026 **Genome-wide association studies (GWAS)**

1027 Case and control definitions for GWAS are listed in **Supplementary Table 3.**
1028 Phenotype exclusions were used only if clinically or biologically justified. We excluded:
1029 1) patients diagnosed with medullary thyroid cancer from thyroid cancer GWAS (if
1030 medullary thyroid cancer data was available; because rare medullary thyroid cancers
1031 are genetically distinct from the common follicular cell-derived thyroid cancers); 2)
1032 thyroid cancer cases from benign nodular goiter GWAS (because all thyroid cancers are
1033 initially diagnosed as thyroid nodules to avoid contamination of benign nodular goiter
1034 cases with malignant tumors), and 3) patients diagnosed with hypothyroidism other than
1035 primary (iatrogenic, congenital, central) from hypothyroidism GWAS.

1036 Each biobank conducted genotyping, imputation, quality control, and genetic
1037 ancestry analysis independently (**Supplementary Table 1**), except for the All of Us
1038 Research Program Biobank, where a custom pipeline was designed to use whole
1039 genome sequencing data and maximize variant overlap with other Biobanks.

1040 GWAS analyses were run using either linear mixed models (SAIGE) [54] or whole
1041 genome regression (REGENIE)[55], adjusted for case-control imbalances using
1042 saddlepoint approximation or Firth's logistic regression. The biobanks were instructed to
1043 use age, sex, up to 20 first principal components, and biobank-specific variables (such
1044 as genotyping batches and recruiting centers) as covariates.

1045 In addition to multi-ancestry analyses, GWAS stratified by genetic ancestry were
1046 performed when the case counts permitted. **Supplementary Table 4** lists case and
1047 control counts, Summix2 [56] population structure estimates, and quality control metrics
1048 calculated with the covariate-adjusted LD-score regression[26] for 198 GWAS analyses
1049 integrated into the meta-analyses.

1050 **Genome-wide association studies in the All of Us (All of Us) Research**
1051 **Program Biobank**

1052 We used All of Us whole genome sequencing v7 (WGS) data (245,388 WGS) to
1053 produce a genetic dataset that maximizes variant overlap with the analyses performed
1054 in the other biobanks (**Extended Data Figures 7 and 8**). An inclusive list of single-
1055 nucleotide polymorphisms (SNPs) and indels from GWAS analyses was compiled and
1056 supplemented with variants from the Polygenic Score Catalog (reported as of February
1057 2024). This list contained ~ 147 million SNPs and indels.

1058 WGS variant level QC was performed by All of Us as reported in the All of Us
1059 Research Program Genomic Research Data Quality Report [59]. In addition, we filtered
1060 the dataset to a maximal set of unrelated samples estimated from kinship scores and
1061 only included individuals with electronic health records and/or survey data for phenotype
1062 definitions (193,429 WGS).

1063 We developed a Hail Python pipeline that extracts variants of interest from the
1064 AoU variant dataset (VDS) (https://hail.is/docs/0.2/vds/index.html). The code is publicly
1065 available in the GitHub repository for this study (https://github.com/pozdeyevlab/vds-
1066 filter/tree/main). The resulting BGEN dataset contained ~118 million directly genotyped
1067 variants (a significant decrease from 972 million variants in VDS), permitting GWAS
1068 using REGENIE version 3.2.4.

1069 **Post-GWAS quality control**

1070 The post-GWAS quality control workflow diagram is shown in **Extended Data**
1071 **Figure 9**. All GWAS summary data were harmonized to gnomAD (v4.1.0) (GRCh38
1072 human genome reference)[28].

1073 Each GWAS summary data set (Supplementary Table 4) was processed using
1074 the following steps.

1075 1. *Variant-level quality control.* The following variants were removed from the
1076 GWAS summary data:
1077 a. Variants containing alleles with characters other than A,T,C, or G.
1078 b. Variants with a p-value equal to zero, effect size (β) or standard error ≥
1079 1e6 or ≤-1e6, and variants with an imputation score < 0.3.
1080 c. Variants with AF < 0.0005 or > 0.9995.
1081 d. Variants with allele count < 20.
1082 e. The variants were aligned to gnomAD (v4.1.0) reference . Ancestry-
1083 specific gnomAD allele frequencies (AFs) were used for single-ancestry
1084 GWAS. Both palindromic and non-palindromic variants were tested for
1085 exact and inverse alignments. Palindromic variants were removed due to
1086 potential strand flip if they met any of the following criteria: the fold
1087 difference between gnomAD AF and GWAS AF > 2; or GWAS AF > 0.4

and < 0.6; or GWAS AF < 0.4 and gnomAD AF > 0.6; or GWAS AF >0.4 and gnomAD AF < 0.6.

f. Variants flagged as low-quality by gnomAD.

g. Variants with Mahalanobis distance between gnomAD AF and harmonized GWAS AF of > 3 standard deviations from the mean.

2. <u>*GWAS summary level data quality control*</u>

a. *Population structure.* Summix2[56] was used to estimate the population structure from the GWAS summary data. We used a random set of 10,000 variants from chromosome 21 and reference allele frequencies for AFR, AMR, EAS, NFE, MID and SAS genetic ancestry groups from gnomAD (v4.1.0). The results from five Summix2 runs, each using a different random set of reference variants, were averaged. We compared GWAS-derived Summix2 population proportion estimates to those published by MVP , CCPM and AoU and found almost perfect agreement (**Extended Data Figure 10**, r^2 = 0.999, p-value = 1.96e-22).
Single-ancestry GWAS summary data analysis showed good agreement between the ancestry reported by the biobank and Summix2 estimate (median fraction of target ancestry is 0.88-0.97).

b. *Covariate-adjusted LD score regression* (cov-LDSC) [26] was used to evaluate for confounding in GWAS summary data, calculate the heritability of phenotypes, and estimate the genetic correlation between thyroid diseases. For each major continental ancestry, we generated a custom reference panel of 5,000 WGS from the AoU Biobank. For multi-ancestry GWAS, we used ancestry proportions calculated with Summix2 (**Supplementary Table 4**). Samples, regions and variants that met at least one of the following criteria were removed: 1) missingness of > 0.1; 2) closely related individuals (*plink king* cutoff of 0.0884); 3) Hardy-Weinberg equilibrium exact test p-value < 1e-6; 4) minor allele frequency of < 0.01, and 4) genomic regions with high linkage disequilibrium (LD). Genetic principal components were calculated using plink2. Ten principal components and a window of 20 cM were used to calculate covariate-adjusted LD scores and estimate LD score regression intercept (**Supplementary Table 4**), heritability (**Supplementary Table 6**), and genetic correlations (**Supplementary Table 7**) [26].

GWAS meta-analysis

Fixed inverse variance weighted meta-analysis was run using METAL software.[62] Individual GWAS summary data with cov-LDSC y-axis intercepts significantly deviating from one were adjusted before meta-analysis.

Post-meta-analysis quality control, variant annotation and classification

To minimize false positive hits introduced by confounding within a single large biobank, only variants present in at least four input GWAS datasets were considered in

the downstream analysis. If three or fewer datasets were available for the ancestry-stratified meta-analysis, then the threshold was set to two. Cochran's Q p-values were calculated to assess heterogeneity across datasets.

We used hg38 human genome reference throughout the study. A genomic position-based algorithm similar to the one published in the GBMI flagship paper[21] was used to define a set of non-overlapping genome-wide significant (p-value ≤ 5e-8) variants. The algorithm ensures that there is at least 500 kb between significant variants in two neighboring loci. Lead variants were mapped to the nearest gene and annotated using ANNOVAR (version date June 7, 2020)[63]. A locus was considered novel if no variants for the corresponding phenotype were reported within ±500 kb in the GWAS catalog (as of April 2024, v.1.0.2) [29]. Otherwise, the variant was labeled as previously discovered.

Heritability estimation and genetic correlation analysis.

We used cov-LDSC (V. 1.0.0) [26] with a custom population structure-matched linkage disequilibrium (LD) reference panel (see the section on GWAS summary level-data quality control) to calculate SNP-based heritability (h^2_{SNP}). Observed-scale heritability estimates and the corresponding SEs were converted to liability scale using phenotype population prevalence calculated in the All of Us v7 dataset (**Supplementary Table 6)**. Similarly, pairwise genetic correlations between the five thyroid phenotypes (**Supplementary Table 7**) were calculated using cov-LDSC with a custom LD-score reference panel.

Polygenic Risk Score Calculation and Evaluation

To calculate and evaluate polygenic risk scores (PRS), we performed a leave-CCPM-biobank-out GWAS meta-analysis. We used similar data derived from the GBMI phase I thyroid cancer meta-analysis [21] in comparison. All PRSs in this study were tested on the current out-of-sample CCPM dataset (n = 94,651). This approach minimizes inflation of PRS performance due to overfitting. Adjusted PRS (covariates of age, sex and ten genetic principal components) were cross-validated (5-fold).

Our PRS was calculated as a weighted sum of independent genome-wide significant risk alleles. For the clinically relevant use case of distinguishing thyroid cancer from benign nodular goiter, we defined $PRS_{ThC\ vs.\ BNG}$ as the difference between PRS for thyroid cancer ($PRS_{ThC\ vs.\ All}$) and PRS for benign nodular goiter ($PRS_{BNG\ vs.\ All}$): $PRS_{ThC\ vs.\ BNG} = PRS_{ThC\ vs.\ All} - PRS_{BNG\ vs.\ All}$. To compare with modern penalized methods, genome-wide PRSs were derived using the Bayesian regression framework within PRS-CS (version date May 14, 2024) [40]. PRS-CS was run with a global shrinkage parameter of 1e-4, and all other fields were set as default.

1166 PRS performance predicting binary phenotypes was assessed using the area
1167 under the receiver operating characteristic curve (AUC). AUCs were compared with
1168 DeLong's test for significant differences.

1169 **Transcriptome-wide association study (TWAS)**

1170 We performed *cis*-acting expression quantitative trait loci (*cis*-eQTL) TWAS using
1171 FUSION [30]. FUSION was run on multi-ancestry and European meta-analysis summary
1172 data, 1000 genomes LD reference data, and all samples thyroid expression reference
1173 weights precomputed from GTEx v8 [33] (http://gusevlab.org/projects/fusion/).

1174 To replicate our findings in FUSION, we also employed the Summary-PrediXcan
1175 (S-PrediXcan) [32] to derive gene-level association results from GWAS summary statistics
1176 and GTEx v8 [33] as the reference set. GWAS meta-analysis summary data were
1177 harmonized and imputed as described previously
1178 (https://github.com/hakyimlab/summary-gwas-imputation). An imputed GWAS was used
1179 to generate gene-trait associations in thyroid gland tissue.
1180

1181 **Candidate gene expression and pathway analysis.**

1182 We studied gene expression of candidate genes linked to significant genetic
1183 associations discovered in the GWAS meta-analysis by ANNOVAR. In addition, we used
1184 significant (at a Bonferroni-corrected p-value threshold) protein-coding genes identified
1185 by *cis*-eQTLs TWAS. Intergenic variants that could not be unambiguously attributed to
1186 the expressed gene were not included.

1187 mRNA expression of thyroid cancer-associated genes was compared in 20
1188 human tissues using the National Center for Biotechnology Information Gene database
1189 (https://www.ncbi.nlm.nih.gov/gene [34]). mRNA expression was analyzed in thyroid
1190 cancers from the Cancer Genome Atlas study [35] and the ORIEN AVATAR Program
1191 (https://www.oriencancer.org/research-programs). Gene expression was compared by
1192 age at diagnosis, cancer stage, somatic mutation status and other tumor features
1193 (**Supplementary Table 10**).

1194 Reactome and KEGG pathway analyses were performed on all significant genes
1195 combined from FUSION and ANNOVAR using *ReactomePA* (v.1.16.2) [38] and
1196 *clusterProfiler* packages in R 4.4 with the default Benjamini-Hochberg adjustment for
1197 multiple hypothesis testing.

1198

1199 **Table 1. Independent genetic loci associated with thyroid diseases.**

	Variant type, N		Variant location, N									
Phenotype	Known	Novel	Exonic	3'UTR	5'UTR	Upstream	Downstream	ncRNA exonic	Intronic	ncRNA intronic	Intergenic	Total
Thyroid cancer	26	61	6	3		2	2		34	11	29	87
Benign nodular goiter	16	84	4	1		0	1	1	52	10	31	100
Graves' disease	30	70	6	1			4		52	6	31	100
Hypothyroidism	120	271	26	9	7	10	5	3	194	30	107	391
Lymphocytic thyroiditis	43	15	4	3		2	3		24	5	17	58
Total	**235**	**501**	**46**	**17**	**7**	**14**	**15**	**4**	**356**	**62**	**215**	**736**

1200 Counts of variants (N) by type and location are shown. Variant type was determined based on genetic associations reported in the NHGRI-
1201 EBI GWAS catalog (April 2024, version 1.0.2). The variant location was extracted from the ANNOVAR annotation.

1202

1 Siegel, R. L., Giaquinto, A. N. & Jemal, A. Cancer statistics, 2024. *CA Cancer J Clin* **74**, 12-49 (2024). https://doi.org/10.3322/caac.21820

2 Paschou, S. A. *et al.* Thyroid disorders and cardiovascular manifestations: an update. *Endocrine* **75**, 672-683 (2022). https://doi.org/10.1007/s12020-022-02982-4

3 Czene, K., Lichtenstein, P. & Hemminki, K. Environmental and heritable causes of cancer among 9.6 million individuals in the Swedish Family-Cancer Database. *Int. J. Cancer* **99**, 260-266 (2002). https://doi.org/10.1002/ijc.10332

4 Lin, H. T. *et al.* Familial Aggregation and Heritability of Nonmedullary Thyroid Cancer in an Asian Population: A Nationwide Cohort Study. *J. Clin. Endocrinol. Metab.* **105** (2020). https://doi.org/10.1210/clinem/dgaa191

5 Iglesias, M. L. *et al.* Radiation exposure and thyroid cancer: a review. *Arch Endocrinol Metab* **61**, 180-187 (2017). https://doi.org/10.1590/2359-3997000000257

6 Rashkin, S. R. *et al.* Pan-cancer study detects genetic risk variants and shared genetic basis in two large cohorts. *Nat Commun* **11**, 4423 (2020). https://doi.org/10.1038/s41467-020-18246-6

7 Brix, T. H. & Hegedüs, L. Twin studies as a model for exploring the aetiology of autoimmune thyroid disease. *Clin Endocrinol (Oxf)* **76**, 457-464 (2012). https://doi.org/10.1111/j.1365-2265.2011.04318.x

8 Dean, D. S. & Gharib, H. Epidemiology of thyroid nodules. *Best Pract Res Clin Endocrinol Metab* **22**, 901-911 (2008). https://doi.org/10.1016/j.beem.2008.09.019

9 Durante, C. *et al.* The Diagnosis and Management of Thyroid Nodules: A Review. *Jama* **319**, 914-924 (2018). https://doi.org/10.1001/jama.2018.0898

10 Tessler, F. N. *et al.* ACR Thyroid Imaging, Reporting and Data System (TI-RADS): White Paper of the ACR TI-RADS Committee. *J. Am. Coll. Radiol.* **14**, 587-595 (2017). https://doi.org/10.1016/j.jacr.2017.01.046

11 Dean, D. S. & Gharib, H. in *Endotext* (eds K. R. Feingold *et al.*) (MDText.com, Inc., 2015).

12 Deaver, K. E., Haugen, B. R., Pozdeyev, N. & Marshall, C. B. Outcomes of Bethesda categories III and IV thyroid nodules over 5 years and performance of the Afirma gene expression classifier: A single-institution study. *Clin Endocrinol (Oxf)* **89**, 226-232 (2018). https://doi.org/10.1111/cen.13747

13 Cawood, T. J. *et al.* TIRADS Management Guidelines in the Investigation of Thyroid Nodules; Illustrating the Concerns, Costs, and Performance. *J Endocr Soc* **4**, bvaa031 (2020). https://doi.org/10.1210/jendso/bvaa031

14 Pozdeyev, N. *et al.* Thyroid Cancer Polygenic Risk Score Improves Classification of Thyroid Nodules as Benign or Malignant. *J. Clin. Endocrinol. Metab.* **109**, 402-412 (2024). https://doi.org/10.1210/clinem/dgad530

15 Gudmundsson, J. *et al.* Common variants on 9q22.33 and 14q13.3 predispose to thyroid cancer in European populations. *Nat. Genet.* **41**, 460-464 (2009). https://doi.org/10.1038/ng.339

16 Gudmundsson, J. *et al.* A genome-wide association study yields five novel thyroid cancer risk loci. *Nat Commun* **8**, 14517 (2017). https://doi.org/10.1038/ncomms14517

17 Son, H.-Y. *et al.* Genome-wide association and expression quantitative trait loci studies identify multiple susceptibility loci for thyroid cancer. *Nature Communications* **8**, 15966 (2017). https://doi.org/10.1038/ncomms15966

18 Köhler, A. *et al.* Genome-wide association study on differentiated thyroid cancer. *J. Clin. Endocrinol. Metab.* **98**, E1674-1681 (2013). https://doi.org/10.1210/jc.2013-1941

19 Mancikova, V. *et al.* Thyroid cancer GWAS identifies 10q26.12 and 6q14.1 as novel susceptibility loci and reveals genetic heterogeneity among populations. *Int. J. Cancer* **137**, 1870-1878 (2015). https://doi.org/10.1002/ijc.29557
20 Truong, T. *et al.* Multiethnic genome-wide association study of differentiated thyroid cancer in the EPITHYR consortium. *Int. J. Cancer* **148**, 2935-2946 (2021). https://doi.org/10.1002/ijc.33488
21 Zhou, W. *et al.* Global Biobank Meta-analysis Initiative: Powering genetic discovery across human disease. *Cell Genom* **2**, 100192 (2022). https://doi.org/10.1016/j.xgen.2022.100192
22 Reeve, M. *et al.* Autoimmune hypothyroidism GWAS reveals independent autoimmune and thyroid-specific contributions and an inverse relation with cancer risk. *Res Sq* (2024). https://doi.org/10.21203/rs.3.rs-4626646/v1
23 Williams, A. T. *et al.* Genome-wide association study of thyroid-stimulating hormone highlights new genes, pathways and associations with thyroid disease. *Nat Commun* **14**, 6713 (2023). https://doi.org/10.1038/s41467-023-42284-5
24 Kurki, M. I. *et al.* FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* **613**, 508-518 (2023). https://doi.org/10.1038/s41586-022-05473-8
25 Verma, A. *et al.* Diversity and scale: Genetic architecture of 2068 traits in the VA Million Veteran Program. *Science* **385**, eadj1182 (2024). https://doi.org/10.1126/science.adj1182
26 Luo, Y. *et al.* Estimating heritability and its enrichment in tissue-specific gene sets in admixed populations. *Hum. Mol. Genet.* **30**, 1521-1534 (2021). https://doi.org/10.1093/hmg/ddab130
27 Stoneman, H. R. *et al.* Characterizing substructure via mixture modeling in large-scale genetic summary statistics. *Am. J. Hum. Genet.* **112**, 235-253 (2025). https://doi.org/10.1016/j.ajhg.2024.12.007
28 Chen, S. *et al.* A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* **625**, 92-100 (2024). https://doi.org/10.1038/s41586-023-06045-0
29 Sollis, E. *et al.* The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res.* **51**, D977-d985 (2023). https://doi.org/10.1093/nar/gkac1010
30 Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245-252 (2016). https://doi.org/10.1038/ng.3506
31 Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**, 1091-1098 (2015). https://doi.org/10.1038/ng.3367
32 Barbeira, A. N. *et al.* Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat Commun* **9**, 1825 (2018). https://doi.org/10.1038/s41467-018-03621-1
33 The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318-1330 (2020). https://doi.org/10.1126/science.aaz1776
34 Duff, M. O. *et al.* Genome-wide identification of zero nucleotide recursive splicing in Drosophila. *Nature* **521**, 376-379 (2015). https://doi.org/10.1038/nature14475
35 Cancer Genome Atlas Research, N. Integrated genomic characterization of papillary thyroid carcinoma. *Cell* **159**, 676-690 (2014). https://doi.org/10.1016/j.cell.2014.09.050
36 Vinagre, J. *et al.* Frequency of TERT promoter mutations in human cancers. *Nat Commun* **4**, 2185 (2013). https://doi.org/10.1038/ncomms3185
37 Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27-30 (2000).
38 Yu, G. & He, Q. Y. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Mol Biosyst* **12**, 477-479 (2016). https://doi.org/10.1039/c5mb00663e

39 Stassi, G. & De Maria, R. Autoimmune thyroid disease: new models of cell death in autoimmunity. *Nat. Rev. Immunol.* **2**, 195-204 (2002). https://doi.org/10.1038/nri750
40 Ge, T., Chen, C. Y., Ni, Y., Feng, Y. A. & Smoller, J. W. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat Commun* **10**, 1776 (2019). https://doi.org/10.1038/s41467-019-09718-5
41 Mulder, J. E. Thyroid disease in women. *Med Clin North Am* **82**, 103-125 (1998). https://doi.org/10.1016/s0025-7125(05)70596-4
42 Canaris, G. J., Manowitz, N. R., Mayor, G. & Ridgway, E. C. The Colorado thyroid disease prevalence study. *Arch Intern Med* **160**, 526-534 (2000). https://doi.org/10.1001/archinte.160.4.526
43 Haugen, B. R. *et al.* 2015 American Thyroid Association Management Guidelines for Adult Patients with Thyroid Nodules and Differentiated Thyroid Cancer: The American Thyroid Association Guidelines Task Force on Thyroid Nodules and Differentiated Thyroid Cancer. *Thyroid* **26**, 1-133 (2016). https://doi.org/10.1089/thy.2015.0020
44 McDermott, M. T. In the clinic. Hypothyroidism. *Ann Intern Med* **151**, Itc61 (2009). https://doi.org/10.7326/0003-4819-151-11-200912010-01006
45 Thomsen, H. *et al.* Familial risks between Graves disease and Hashimoto thyroiditis and other autoimmune diseases in the population of Sweden. *J Transl Autoimmun* **3**, 100058 (2020). https://doi.org/10.1016/j.jtauto.2020.100058
46 Goel, R. *et al.* A Signaling Network of Thyroid-Stimulating Hormone. *J Proteomics Bioinform* **4** (2011). https://doi.org/10.4172/jpb.1000195
47 Chen, Y. K. *et al.* Cancer risk in patients with Graves' disease: a nationwide cohort study. *Thyroid* **23**, 879-884 (2013). https://doi.org/10.1089/thy.2012.0568
48 DeBoy, E. A. *et al.* Telomere-lengthening germline variants predispose to a syndromic papillary thyroid cancer subtype. *Am. J. Hum. Genet.* **111**, 1114-1124 (2024). https://doi.org/10.1016/j.ajhg.2024.04.006
49 Liyanarachchi, S. *et al.* Assessing thyroid cancer risk using polygenic risk scores. *Proc Natl Acad Sci U S A* **117**, 5997-6002 (2020). https://doi.org/10.1073/pnas.1919976117
50 Bibbins-Domingo, K. *et al.* Screening for Thyroid Cancer: US Preventive Services Task Force Recommendation Statement. *Jama* **317**, 1882-1887 (2017). https://doi.org/10.1001/jama.2017.4011
51 Nam-Goong, I. S. *et al.* Ultrasonography-guided fine-needle aspiration of thyroid incidentaloma: correlation with pathological findings. *Clin Endocrinol (Oxf)* **60**, 21-28 (2004). https://doi.org/10.1046/j.1365-2265.2003.01912.x
52 Reuters, K. B. *et al.* Bethesda Classification and Cytohistological Correlation of Thyroid Nodules in a Brazilian Thyroid Disease Center. *Eur Thyroid J* **7**, 133-138 (2018). https://doi.org/10.1159/000488104
53 Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584-591 (2019). https://doi.org/10.1038/s41588-019-0379-x
54 Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* **50**, 1335-1341 (2018). https://doi.org/10.1038/s41588-018-0184-y
55 Mbatchou, J. *et al.* Computationally efficient whole-genome regression for quantitative and binary traits. *Nat. Genet.* (2021). https://doi.org/10.1038/s41588-021-00870-7
56 Stoneman, H. R. *et al.* Characterizing substructure via mixture modeling of genetic similarity in large-scale summary statistics. *bioRxiv*, 2024.2001.2029.577805 (2024). https://doi.org/10.1101/2024.01.29.577805

57 Lambert, S. A. *et al.* The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation. *Nat. Genet.* **53**, 420-425 (2021). https://doi.org/10.1038/s41588-021-00783-5

58 Lambert, S. A. *et al.* Enhancing the Polygenic Score Catalog with tools for score calculation and ancestry normalization. *Nat. Genet.* **56**, 1989-1994 (2024). https://doi.org/10.1038/s41588-024-01937-x

59 Genomic data in the All of Us Research Program. *Nature* **627**, 340-346 (2024). https://doi.org/10.1038/s41586-023-06957-x

60 Wiley, L. K. *et al.* Building a vertically integrated genomic learning health system: The biobank at the Colorado Center for Personalized Medicine. *Am. J. Hum. Genet.* **111**, 11-23 (2024). https://doi.org/10.1016/j.ajhg.2023.12.001

61 Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015). https://doi.org/10.1186/s13742-015-0047-8

62 Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190-2191 (2010). https://doi.org/10.1093/bioinformatics/btq340

63 Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research* **38**, e164-e164 (2010). https://doi.org/10.1093/nar/gkq603

64 Xu, S. *et al.* Using clusterProfiler to characterize multiomics data. *Nat Protoc* **19**, 3292-3320 (2024). https://doi.org/10.1038/s41596-024-01020-z

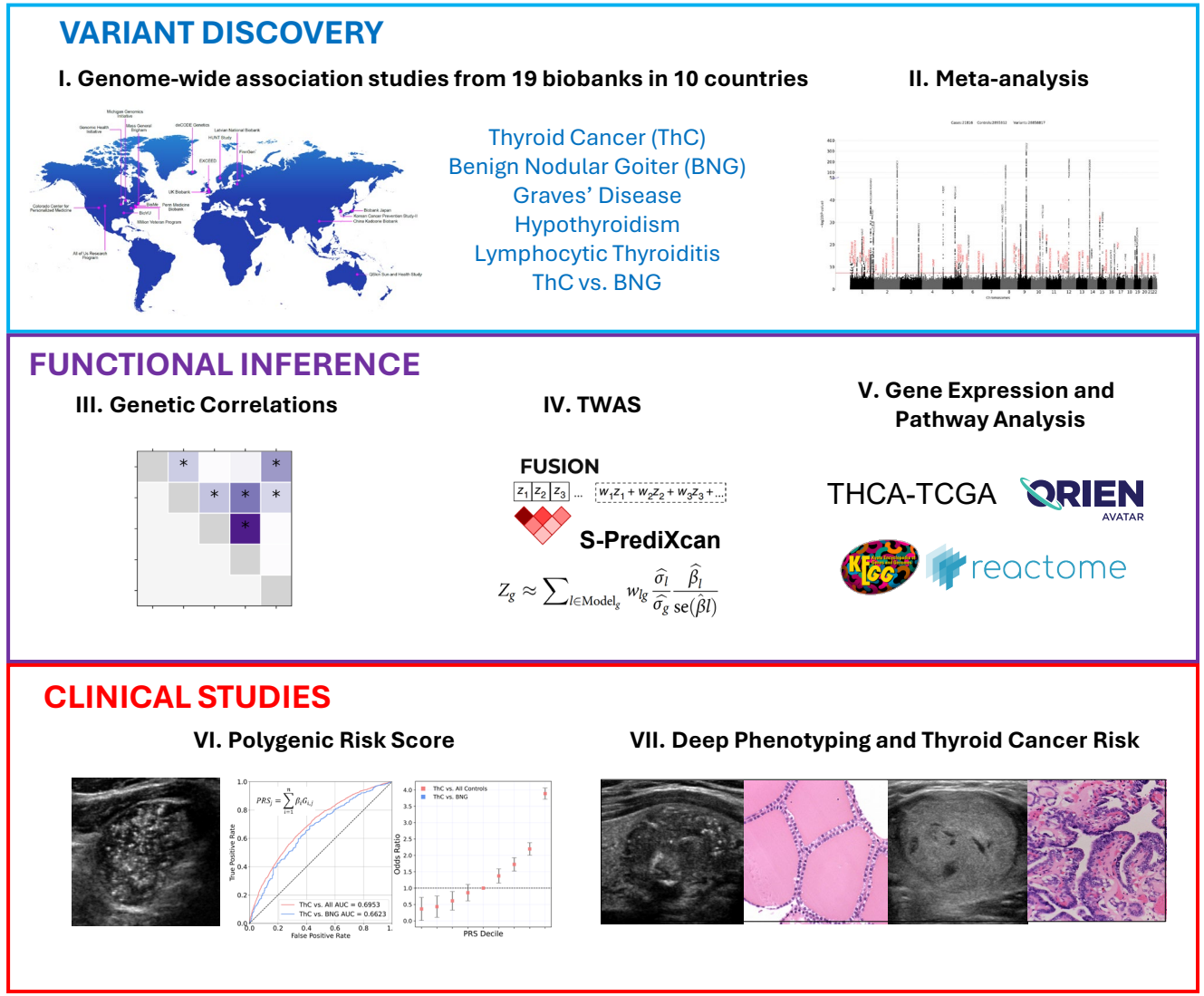


Figure 1. The study design. I. The Virtual Thyroid Biopsy Consortium was formed under the Global Biobank Meta-analysis Initiative. Participating biobanks performed genome-wide association studies (GWAS) for five thyroid diseases and GWAS of thyroid cancer vs. benign nodule goiters (ThC vs. BNG). **II.** Inverse variance-weighted meta-analysis was done after quality-control procedures. Previously known and novel independent genetic associations were identified. Functional inference studies included: **III**. Genetic correlation analysis with covariate-adjusted LD score regression. **IV. T**ranscriptome-wide association studies (FUSION and S-PrediXcan), **V.** Pathway (KEGG and Reactome), and gene-expression (The Cancer Genome Atlas and ORIEN Avatar) analyses. **VI.** Polygenic risk scores were developed for thyroid cancer, benign thyroid diseases, and to distinguish malignant and benign thyroid nodules. **VII.** Polygenic risk scores were tested for association with thyroid diseases and aggressive thyroid cancer features extracted from clinical charts and surgical histopathology reports.

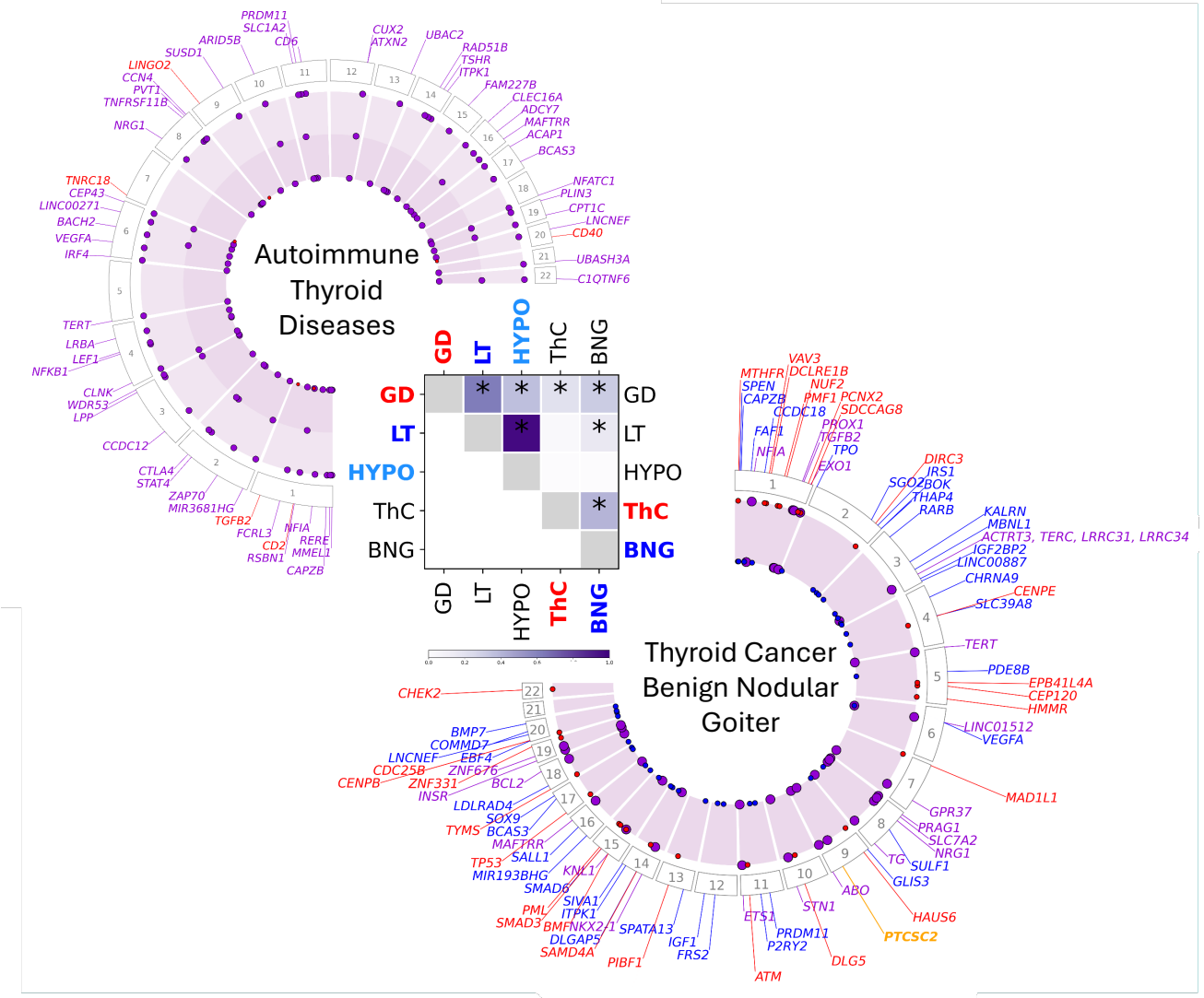
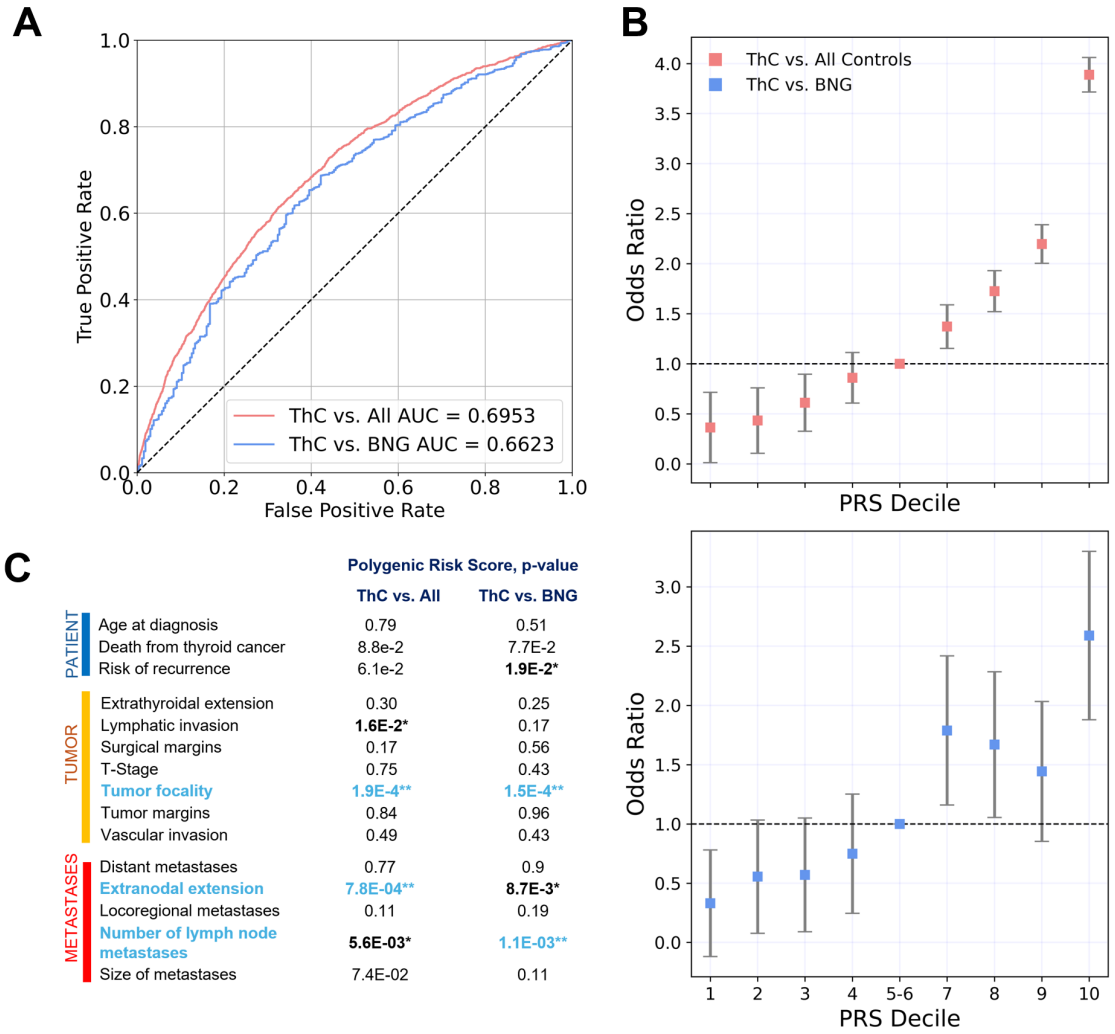


Figure 2. Pleiotropic and phenotype-specific loci associated with thyroid diseases in multi-ancestry GWAS meta-analysis. The heatmap illustrates the genetic correlation (r^2) between thyroid phenotypes, which was estimated using covariate-adjusted LD score regression. The asterisks denote Benjamini-Hochberg false discovery rate (FDR) <0.05. Circular plots highlight loci significantly associated with thyroid cancer and benign nodular goiter (right), and autoimmune thyroid diseases (left). Pleiotropic loci are marked by large purple dots. In the circular plot on the right small dots show loci predominantly associated with thyroid cancer or benign nodular goiter (red or blue respectively). In the circular plot on the left the small red dots depict loci which are significantly associated with Graves' disease, but not lymphocytic thyroiditis or primary hypothyroidism. For clarity, only loci significantly associated with Graves' disease are shown on the autoimmune thyroid diseases circular plot. The *PTCSC2* (right, yellow) is the only locus inversely associated with thyroid cancer and benign nodular goiter.



C

		Polygenic Risk Score, p-value	
		ThC vs. All	ThC vs. BNG
PATIENT	Age at diagnosis	0.79	0.51
	Death from thyroid cancer	8.8e-2	7.7E-2
	Risk of recurrence	6.1e-2	**1.9E-2***
TUMOR	Extrathyroidal extension	0.30	0.25
	Lymphatic invasion	**1.6E-2***	0.17
	Surgical margins	0.17	0.56
	T-Stage	0.75	0.43
	Tumor focality	**1.9E-4****	**1.5E-4****
	Tumor margins	0.84	0.96
	Vascular invasion	0.49	0.43
METASTASES	Distant metastases	0.77	0.9
	Extranodal extension	**7.8E-04****	**8.7E-3***
	Locoregional metastases	0.11	0.19
	Number of lymph node metastases	**5.6E-03***	**1.1E-03****
	Size of metastases	7.4E-02	0.11

Figure 3. Thyroid cancer polygenic risk scores. Two thyroid cancer polygenic risk scores (PRS) were developed: $PRS_{ThC\ vs.\ All}$ to identify individuals at risk in a population and $PRS_{ThC\ vs.\ BNG}$ for clinically relevant task of discriminating malignant and benign thyroid nodules. PRS were tested in the Colorado Center for Personalized Medicine Biobank (CCPM) population (n = 94,651) and CCPM data was not used for PRS development. $PRS_{ThC\ vs.\ All}$ was derived using PRS-CS method. $PRS_{ThC\ vs.\ BNG}$ was calculated using independent significant associations with thyroid cancer and benign nodular goiter. **A.** Receiver operating characteristic curves. **B.** Thyroid cancer risk by PRS decile. **C.** PRS association with features of aggressive thyroid cancer. * - p-value ≤ 0.05. ** - p-value ≤ 1.7e-3 (Bonferroni-corrected significance threshold).

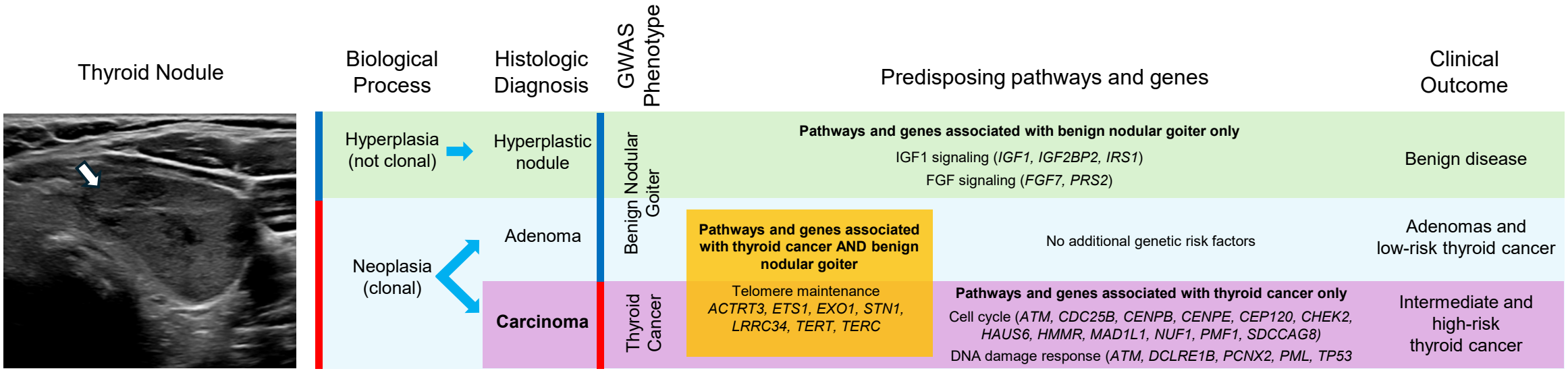
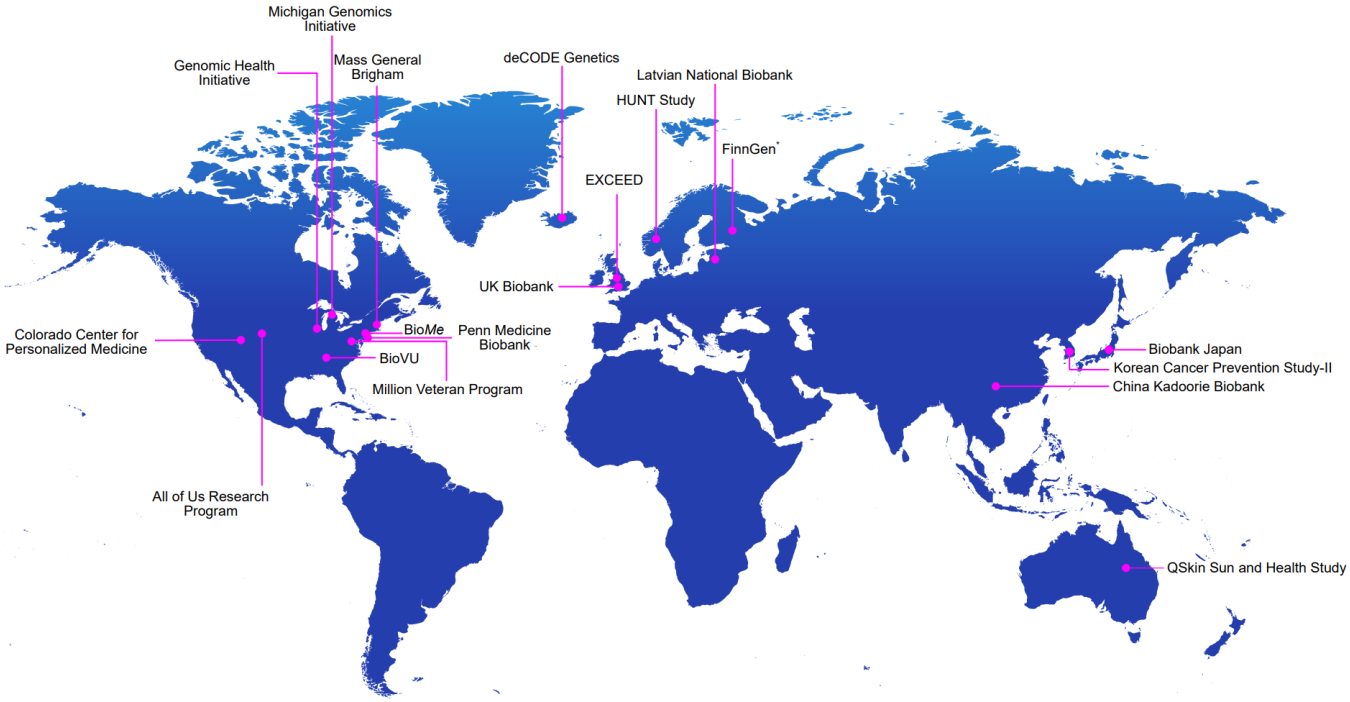
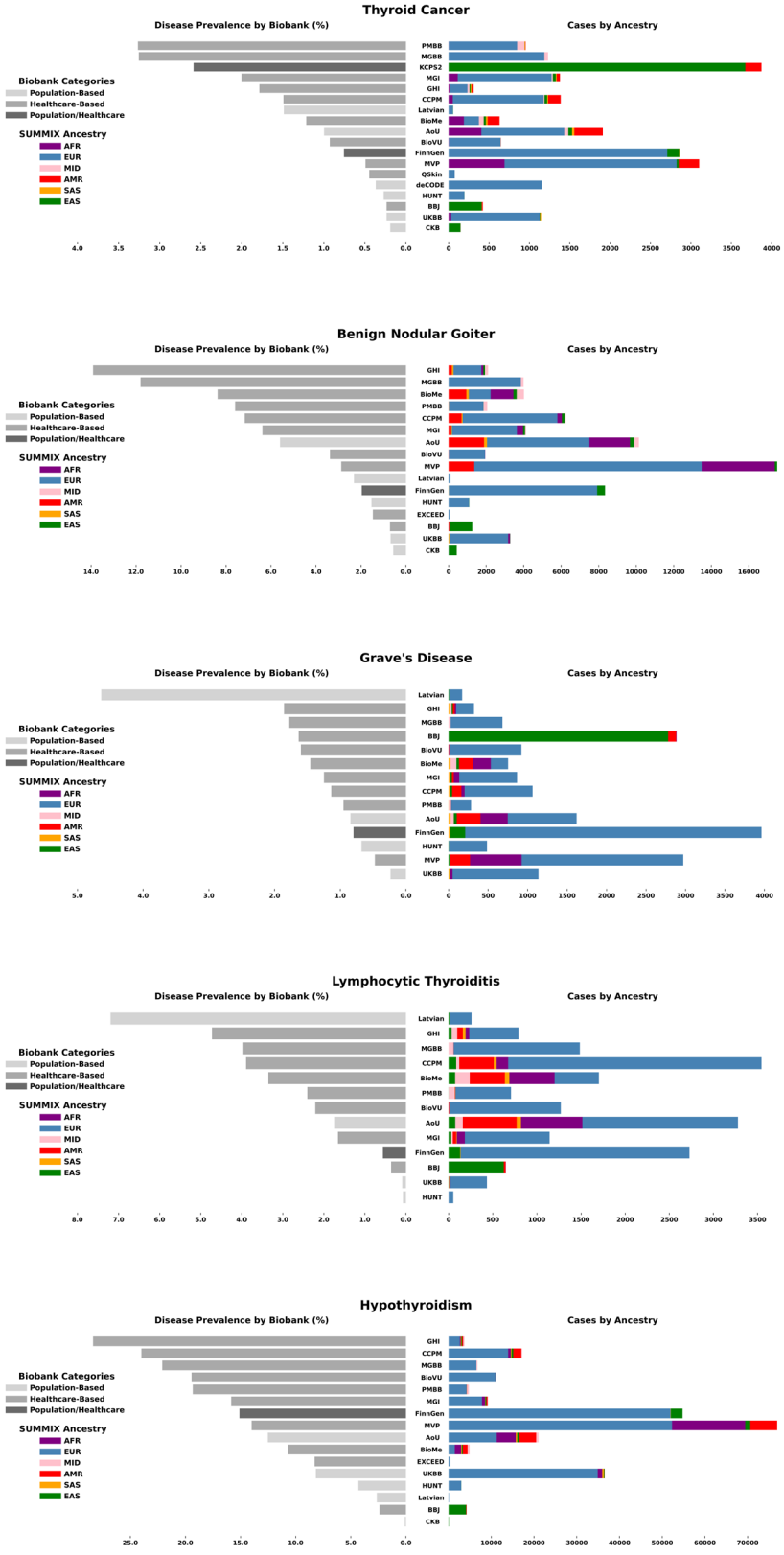


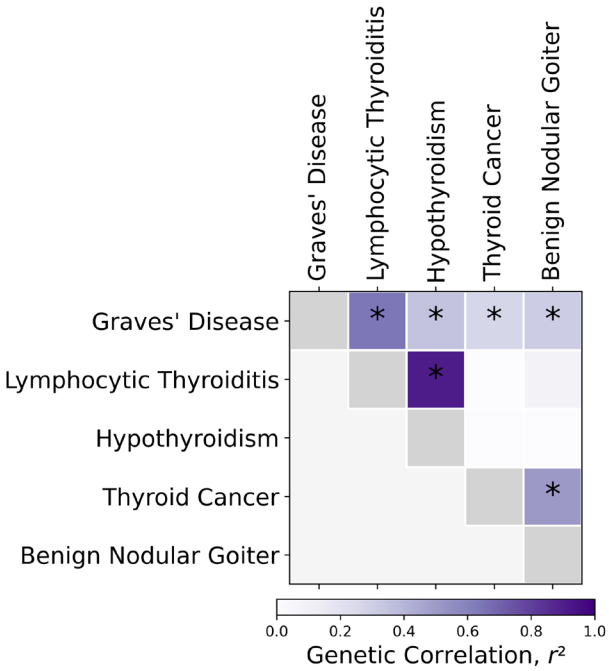
Figure 4. Germline genetic susceptibility to thyroid cancer and benign nodular goiter. We hypothesize that two biological processes with distinct genetic architecture cause thyroid nodules: 1) hyperplasia, a polyclonal follicular cell proliferation with no malignant potential, and 2) neoplasia, a clonal growth driven by somatic genetic alterations. Neoplastic nodules can be benign or malignant, and the mismatch between biological mechanisms (hyperplasia vs. neoplasia) and GWAS phenotype definitions (benign and malignant thyroid nodules) caused apparent genetic pleiotropy. Pathway and genes associated with benign nodular goiter but not thyroid cancer in GWAS meta-analysis (e.g., insulin-like growth factor 1 [IGF1] and fibroblast growth factor [FGF] signaling pathways) predispose to benign nodules that have no malignant potential (green). Pathways and genes associated with both benign nodular goiter and thyroid cancer (e.g., telomere maintenance) predispose to neoplastic thyroid nodules, benign or malignant (orange). In the absence of other genetic risk factors, patients are more likely to develop benign adenomas or low risk thyroid cancers (blue). Alternatively, genetic alterations in cell cycle and DNA damage response genes (purple, associated predominantly with thyroid cancer but nor benign nodular goiter in GWAS meta-analysis) predispose to high-risk thyroid cancer.



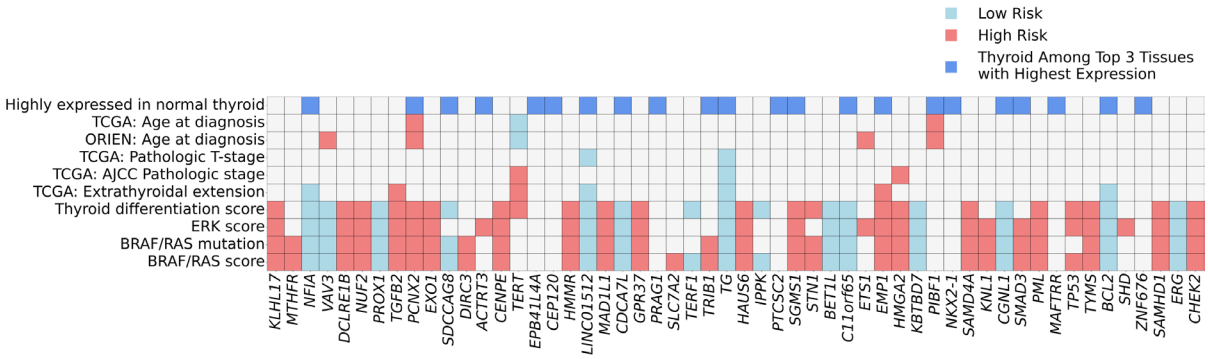
Extended Data Figure 1. Virtual Thyroid Biopsy Consortium. The Consortium aggregated data from 19 biobanks, 10 countries, four continents and ~2.9 million participants.



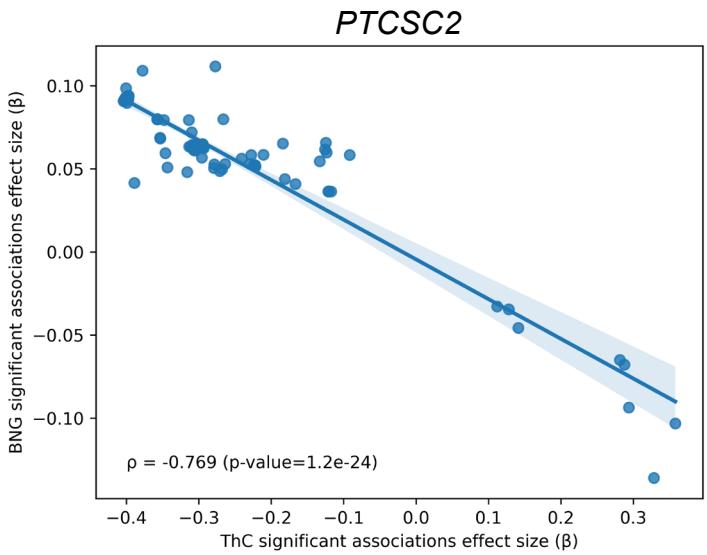
Extended Data Figure 2. Thyroid disease meta-analysis case counts and prevalence across biobanks.



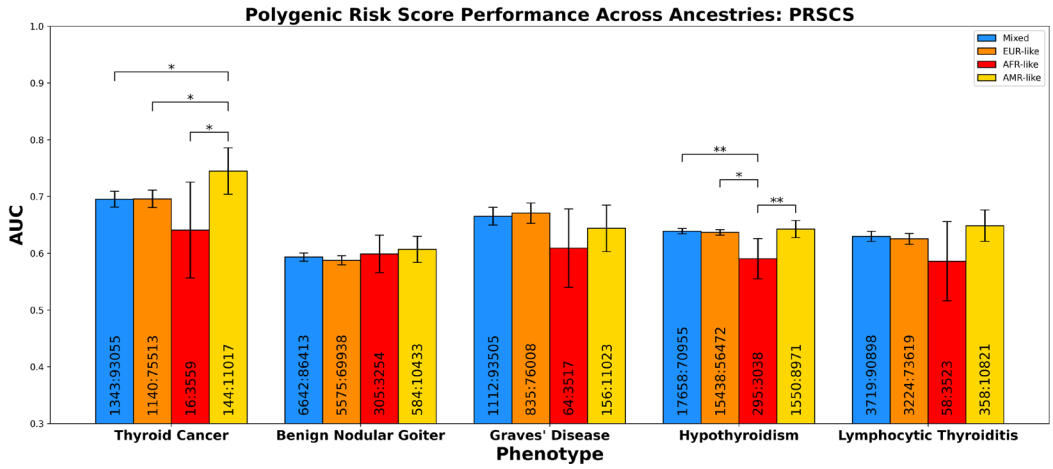
Extended Data Figure 3. Genetic correlation analysis for thyroid diseases in EUR-like GWAS meta-analysis. Genetic correlations were estimated using covariate-adjusted LD score regression. The asterisks denote Benjamini-Hochberg false discovery rate (FDR) <0.05.



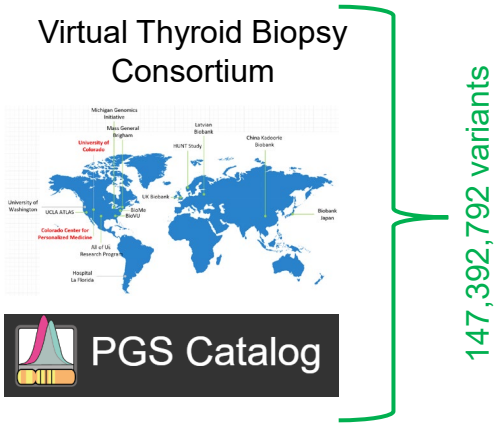
Extended Data Figure 4. mRNA expression of thyroid cancer-associated genes in normal thyroid tissue and thyroid cancer. Genes were identified from ANNOVAR annotations of genome-wide significant variants in thyroid cancer GWAS meta-analysis and FUSION TWAS *cis*-eQTL analysis. Dark blue color shows genes with high expression in normal thyroid tissue (thyroid is among the top three highest expressing tissues in pan-tissue transcriptome analysis from NCBI Gene database https://www.ncbi.nlm.nih.gov/gene). Significant associations (at Bonferroni-corrected p-value ≤ 9.1e-05, red = positive; light blue = negative) of mRNA expression with high-risk thyroid cancer features such as presence of somatic *BRAF* V600E mutation, higher ERK score and lower thyroid differentiation score.



Extended Data Figure 5. Scatterplot of effect sizes of the variants in *PTCSC2* locus significantly (p-value < 5e-8) associated with thyroid cancer and benign nodular goiter. ThC – thyroid cancer. BNG – benign nodular goiter. ρ - Spearman correlation.

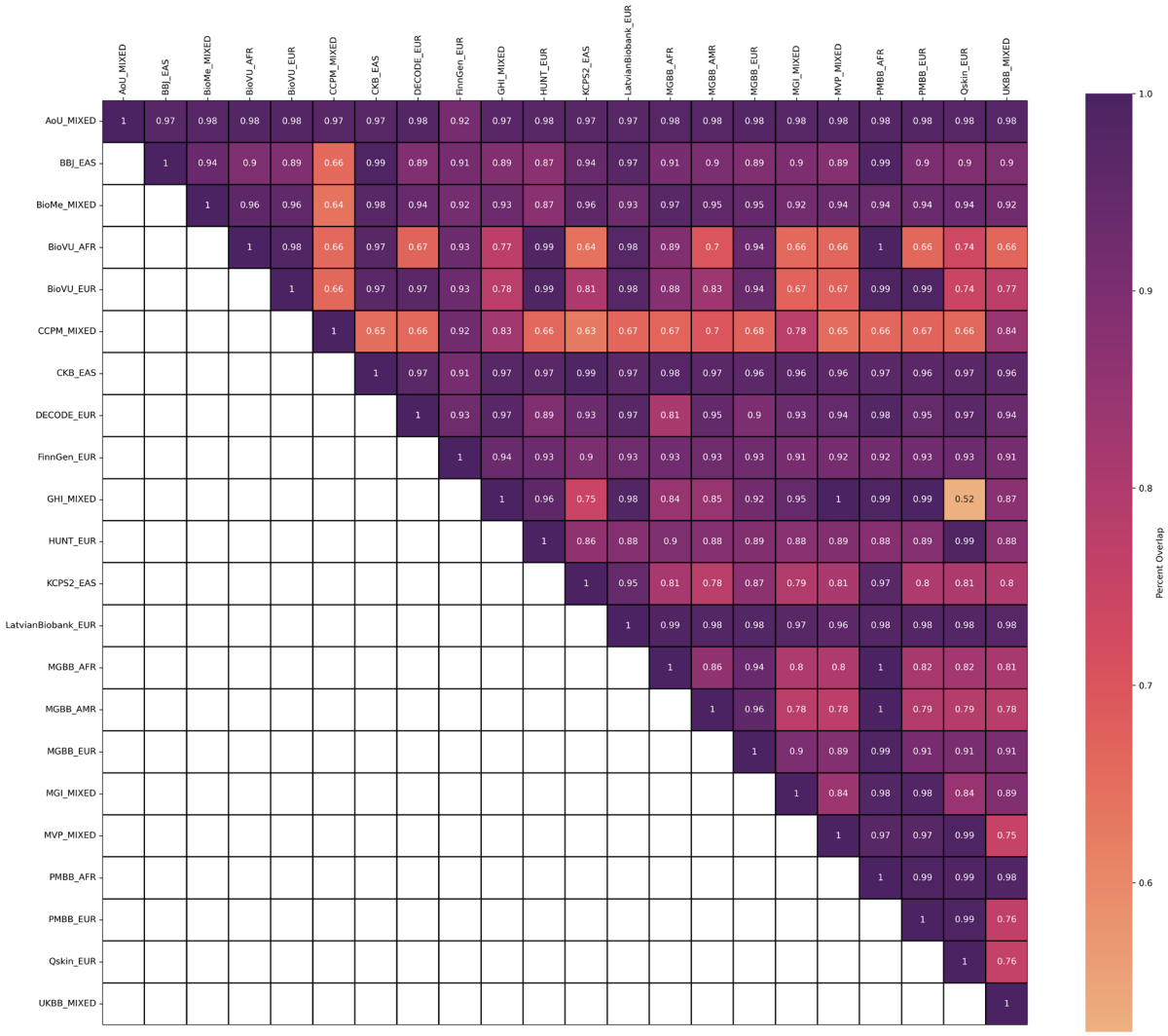


Extended Data Figure 6. Thyroid cancer polygenic risk score ($PRS_{ThC\ vs.\ BNG}$) performance by ancestry. Area under the receiver operating characteristic curve (AUC) and 95% confidence interval are shown for $PRS_{ThC\ vs.\ BNG}$ developed from the leave-CCPM biobank-out thyroid cancer meta-analysis and tested on CCPM biobank participants of various ancestries. Case and control counts are shown within bars. * - p-value ≤ 0.05; ** - p ≤ 0.01.

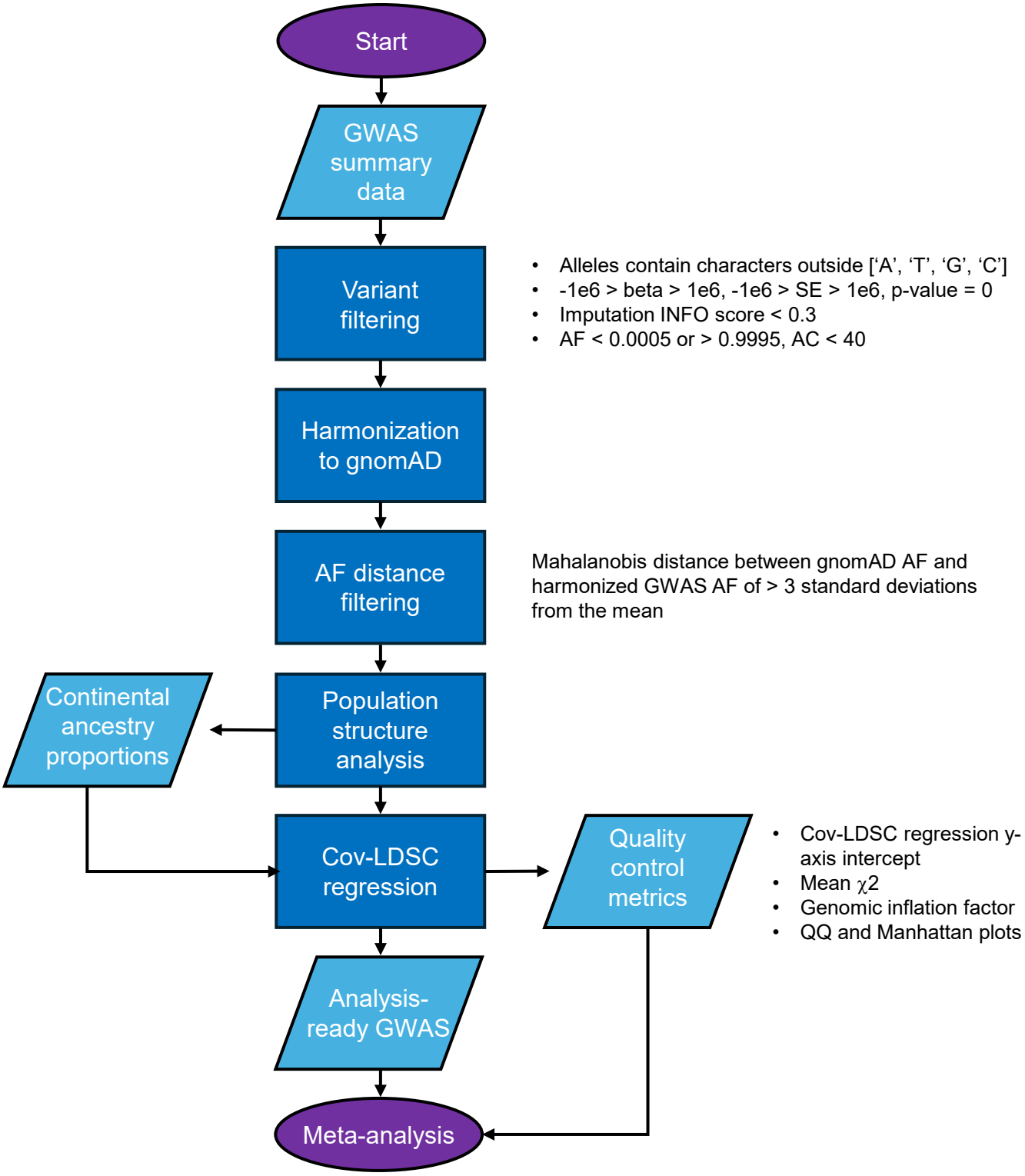


1. Split VDS by chromosome
2. Split multiallelic sites
3. Subset for variants of interest
4. Densify genotypes
5. Annotate rows
6. Drop irrelevant fields
7. Filter variants by allele count (≥20)
8. Add dosages
9. Add rsid information
10. Export BGEN

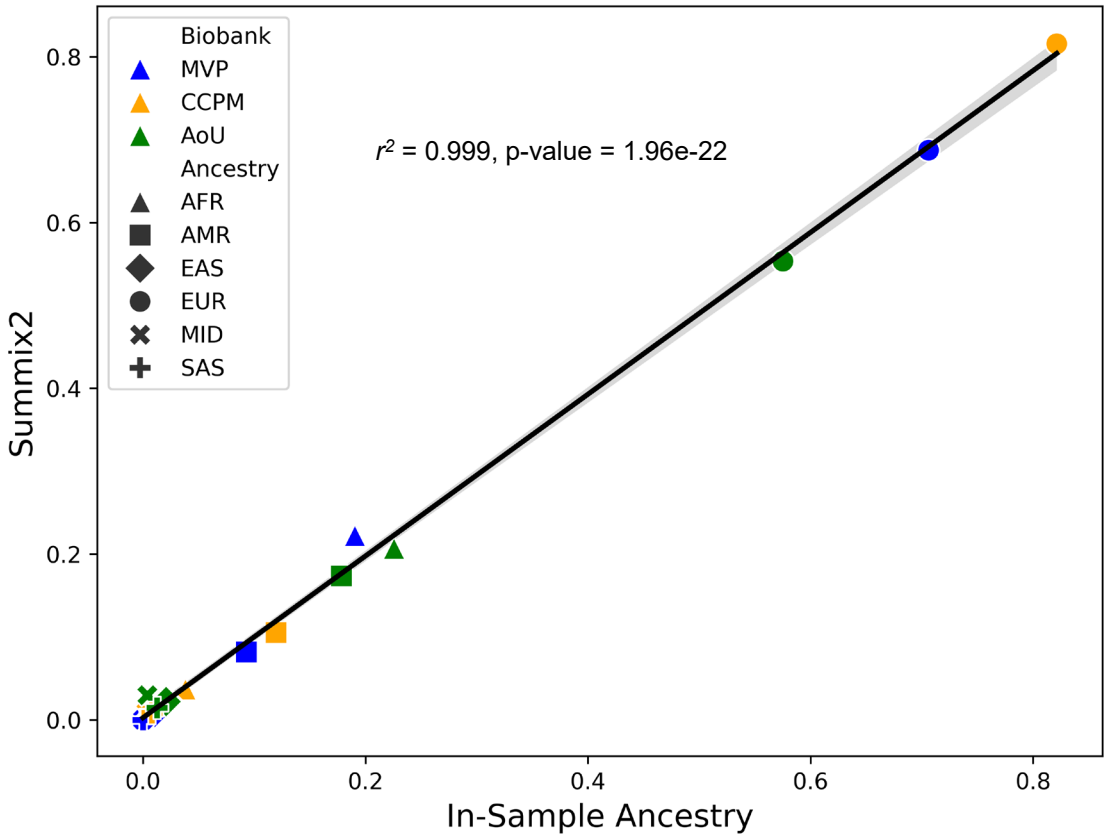
Extended Data Figure 7. Whole genome sequencing data pipeline for the All of Us Research Program data. Variants (SNPs and indels) from participating biobanks' GWAS summary data and PGS Catalog were extracted from the Hail variant dataset v7 object.



Extended Data Figure 8. Variant overlap between GWAS from participating Biobanks. Data is shown as a fraction of variants that are identical by chromosome, position, reference and alternate allele in harmonized GWAS summary data from two biobanks. The All of Us Research Program GWAS (top row) was performed on whole genome sequencing data and was designed to maximize variant overlap with other biobanks.



Extended Data Figure 9. Post-GWAS quality control workflow diagram. AF – allele frequency. AC – allele count, cov-LDSC – covariate-adjusted LD score regression. QQ plot – quantile-quantile plot.



Extended Data Figure 10. Correlation of major continental ancestry fractions estimated by Summix2 (y-axis) and published by the Million Veteran Program, Colorado Center for Personalized Medicine and All of Us Research Program Biobanks (x-axis). Multi-ancestry GWAS summary data was used for this analysis.