Assessing Thyroid cancer risk using polygenic risk scores

Sandya Liyanarachchi[a,1], Julius Gudmundsson[b,1], Egil Ferkingstad[b], Huiling He[a], Jon G. Jonasson[c,d,e], Vinicius Tragante[b,f], Folkert W Asselbergs[f,g,h,i], Li Xu[j], Lambertus A. Kiemeney[k], Romana T. Netea-Maier[l], Jose I. Mayordomo[m], Theo S. Plantinga[n], Hannes Hjartarson[c], Jon Hrafnkelsson[c], Erich M. Sturgis[j], Pamela Brock[o], Fadi Nabhan[p], Gudmar Thorleifsson[b], Matthew D. Ringel[p], Kari Stefansson[b,d,2], and Albert de la Chapelle[a,2]

[a]Human Cancer Genetics Program and Department of Cancer Biology and Genetics, Comprehensive Cancer Center, The Ohio State University, Columbus, Ohio, 43210, USA;

[b]deCODE genetics/Amgen Inc., 101 Reykjavik, Iceland; [c]Landspitali-University Hospital, 101 Reykjavik, Iceland; [d]Faculty of medicine, University of Iceland, 101 Reykjavik, Iceland; [e]The Icelandic Cancer Registry, 105 Reykjavik, Iceland; [f]Department of Cardiology, Division Heart & Lungs, University Medical Center Utrecht, University of Utrecht, 3584 CX Utrecht, The Netherlands; [g]Durrer Center for Cardiovascular Research, Netherlands Heart Institute, 3511 EP Utrecht, The Netherlands; [h]Institute of Cardiovascular Science, Faculty of Population Health Sciences, University College London, WC1E 6BT London, United Kingdom; [i]Farr Institute of Health Informatics Research and Institute of Health Informatics, University College London, NW1 2DA

London, United Kingdom; [j]Department of Head & Neck Surgery and Department of Epidemiology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA; [k]Radboud University Medical Centre, Radboud Institute for Health Sciences, 6500HB Nijmegen, The Netherlands; [l]Division of Endocrinology, Department of Internal Medicine, Radboud University Medical Centre, Radboud Institute for Health Sciences, 6500HB Nijmegen, The Netherlands; [m]University of Colorado Hospital, Aurora, CO 80045, USA; [n]Department of Pathology, Radboud University Medical Center, Radboud Institute for Molecular Life Sciences, 6500HB Nijmegen, The Netherlands; [o]Department of Internal Medicine, Comprehensive Cancer Center, The Ohio State University, Columbus, Ohio, 43210, USA; [p]Division of Endocrinology, Diabetes, and Metabolism, The Ohio State University, Columbus, Ohio, 43210, USA.

[1] S.L. and J.G. contributed equally to this work.

[2] K.S., and A.d.l.C. contributed equally to this work.

To whom correspondence may be addressed:

Albert de la Chapelle, Human Cancer Genetics Program and Department of Cancer Biology and Genetics, Comprehensive Cancer Center, The Ohio State University, Columbus, Ohio, 43210, USA, phone: 614-688-4781, email: albert.delachapelle@osumc.edu

Kari Stefansson, deCODE genetics/Amgen Inc., 101 Reykjavik, Iceland, phone: 354-570-1900, email: kstefans@decode.is

Classification

BIOLOGICAL SCIENCES: Genetics

Keywords: Thyroid cancer, GWAS, Polygenic risk score, Risk prediction

Author Disclosure Statement

The authors from deCODE/AMGEN are employees of deCODE/AMGEN. The remaining authors have no conflicts of interest to declare.

Abstract

Genome-wide association studies (GWAS) have identified at least ten single-nucleotide polymorphisms (SNPs) associated with papillary thyroid cancer (PTC) risk. Most of these SNPs are common variants with small to moderate effect sizes. Here we assessed the combined genetic effects of these variants on PTC risk by using summarized GWAS results to build polygenic risk score (PRS) models in three PTC study groups from Ohio, US (1,544 patients and 1,593 controls), Iceland (723 patients and 129,556 controls), and the United Kingdom (UK) (534 patients and 407,945 controls). A PRS based on the 10 established PTC SNPs showed a stronger predictive power compared with the clinical factors model with a minimum increase of area under the receiver-operating curve of 5.4 percentage points ($P \leq 1.0 \times 10^{-9}$). Adding an extended PRS based on 592,475 common variants did not significantly improve the prediction power compared to the 10 SNPs model, suggesting that most of the remaining undiscovered genetic risk in thyroid cancer is due to rare, moderate to high penetrance variants rather than common low-penetrance variants. Based on 10-SNP PRS, individuals in the top decile group of PRSs have a close to 7-fold risk (95% CI 5.4-8.8) compared to the bottom decile group. In conclusion, polygenic risk scores based on a small number of common germline variants emphasize the importance of heritable low-penetrance markers in PTC.

Significance Statement

Thyroid cancer shows a high degree of heritability in comparison with other cancers. GWASs have identified at least 10 SNPs associated with PTC risk. How these risk factors might help in individualizing the assessment of thyroid cancer risk clinically is unexplored. We present PRS analysis with consistent results in three large cohorts (US, Iceland, and UK). The 10 GWAS SNPs have additive effects on cancer predisposition and the 10-SNP PRS has equally strong risk predictive power compared to a PRS with greater than 500,000 common variants. Our work demonstrates that the 10 low-penetrance variants have the potential to be applied in medicine to improve individualized cancer risk assessment.

Introduction

Recent advances in genetic and genomic research have led to the development of efficient methods to detect and evaluate diagnostic and prognostic factors in the individual patient. In numerous monogenic and congenital disorders, diagnoses can be made based on the occurrence of germline variants. In contrast, polygenic and acquired disorders can be assessed by the study of somatic mutations in the appropriate cell or tissue. Typically, this applies to many cancers where the study of genetic mutations in the tumor itself can be diagnostic, prognostic, or informative for choice of therapy (1-3). The present study investigates methods for thyroid cancer risk assessment at the germline level.

Thyroid cancer is the 9th most common type of cancer in the world, with an annual incidence of over 500,000 cases (over 50,000 of which in the United States) (4,

5). Although surgery and other therapies solve most cases, morbidity among patients is high, and in some cases, the tumors can have more aggressive behavior. Thyroid cancer can be categorized by histology. The medullary type accounts for approximately 5% of all cases arises from parafollicular C-cells of the thyroid. The remaining 95% of all thyroid cancer cases are of the non-medullary type and arise in cells of follicular origin. There are three major histological forms of non-medullary thyroid cancer: papillary (PTC), follicular (FTC) and anaplastic (ATC). PTC alone accounts for ~85% of all thyroid cancers.

The concept of this study has already been used in many investigations dealing with risk assessment in cancers (4-8). Here we are addressing an aspect of this field that has not yet been fully examined, namely the phenotypic effect of low-penetrance mutations and their use in individualized thyroid cancer risk assessment. PTC mutations with high penetrance have been found but account for a very minor part of all PTC. We and others have suggested that common alleles each conferring a slightly increased risk may account for many PTCs and provide an explanation for the high heritability of PTC (9-11). Indeed, many common, low-penetrance SNPs have been found to convey PTC risk (12-21). In this study we analyze the combined genetic effects of 10 well established thyroid cancer risk SNPs, by constructing and evaluating their polygenic risk scores (PRSs). We also checked for any residual predictive power in the remaining genome by generating a genome-wide PRS using 592K tagging SNPs and adding it to the 10-SNP PRS as well as including conventional clinical factors.

Results

Study participants and their demographic characteristics. The results reported here are based on previously published GWAS of thyroid cancer, with populations from Columbus, Ohio and Houston, Texas, in the United States (US), Iceland, the Netherlands and Spain (16). In addition, we have incorporated thyroid cancer GWAS data generated using genotypic information from the UK Biobank (UKB) (22, 23). The addition of UKB results to the meta-analysis of thyroid cancer did not reveal any new genome-wide significantly associated risk variants.

The calculation of the polygenic risk scores (PRS) was done in the three largest sample sets coming from Ohio, Iceland and the UK. The Ohio study group is comprised of 1,544 thyroid cancer samples, and 1,593 controls (Table 1). From Iceland, 723 thyroid cancer samples and 129,556 controls were used, and from the UK we used 534 thyroid cancer cases and 407,945 controls (Table 1).

Polygenic risk score (PRS) analysis in study groups from Ohio, Iceland, and the UK and association with cancer risk. The effect estimates included in the PRS analysis are based on the meta-analysis of thyroid cancer including all study groups listed above. In short, we generated PRSs for the Ohio study group by using effect estimates after excluding all samples from the US from our thyroid cancer GWAS meta-analysis, thereby omitting any potential confounding effects. Similarly, when generating PRS for Icelandic and British individuals, corresponding samples from those study groups were excluded from the meta-analysis (Table 2).

For each individual belonging to the Ohio, Icelandic or UK study groups, a PRS was generated using the published 10 GWAS thyroid cancer risk SNPs (10-SNP PRS; see Table 2) as well as using 592,475 (592K) common SNPs with minor allele frequency > 1% (592K-SNP PRS) . PRS for the 592K common SNPs is estimated based on LDpred method adjusting for GWAS summary statistics for the effects of linkage disequilibrium (24). The risk loci, risk allele frequencies, and the effect estimates of the 10 GWAS SNPs included in the 10-SNP PRS are provided in Table 2. The PRSs of the 10-SNPs and the 592K-SNPs are approximately normally distributed among thyroid cases and controls (*SI Appendix*, Fig S1) and are significantly different between thyroid cancer cases and controls in all three study groups ($p= 2.9\times10^{-58}$, $p= 3.3\times10^{-48}$ in Ohio; $p= 1.3\times10^{-48}$, $p= 4.7\times10^{-33}$ in Iceland; and $p= 1.3\times10^{-25}$, $p= 2.2\times10^{-23}$ in UK, respectively, for the 10-SNP PRS and the 592K-SNP PRS)

PRSs in prediction models. In order to investigate the predictive ability of PRSs, we evaluated prediction models using receiver operating characteristic (ROC) curves. With clinical factors (CF), including: year of birth (YOB), gender, the 10 first principal components (PCs), and familiality (not available for the UKB samples), we obtained an area under the ROC curve (AUC) of 0.585 (95% CI 0.565-0.605) in the Ohio study group. Using the Icelandic- and UKB study groups the results amounted to an AUC of 0.697 (95% CI 0.680-0.714) and 0.629 (95% CI 0.606-0.651), respectively (see Fig 1, and Table 3). By adding the10-SNP PRS to the Ohio model with CF, we obtained a

significantly increased AUC of 0.692 ($p=3.1\times10^{-21}$; Fig 1*A*, Table 3). Similarly, by adding 10-SNP PRS to the model with CF for Iceland and UKB, we obtained significantly increased AUCs of 0.751 ($p=3.0\times10^{-14}$) and 0.694 ($p=1.0\times10^{-09}$), respectively, (Fig 1*B* and 1*C*, Table 3).

We further evaluated the prediction ability after adding the 592K-SNP PRS to the model with 10-SNP PRS and CF (*SI Appendix*, Table S2 for results for individual covariates). For the Ohio and the Icelandic samples AUCs of 0.693 and 0.752, respectively, were obtained, showing only a 0.1 percentage point increase over the 10-SNP PRS model (p=0.34 and p=0.31) (Fig 1*A*, 1*B* and Table 3). In the UKB samples an AUC of 0.697 was obtained, or a non-significant 0.3 percentage point increase (p=0.27) (Fig 1*C* and Table 3). Together, these results demonstrate that no significant improvement was achieved by adding the genome–wide PRS (592K-SNP PRS) to the model.

In a multi-collinearity analysis, we observed variance inflation factors (VIFs) of 2.32 and 2.32, and of 2.04 and 2.04 in the Ohio and UK study groups, respectively and for the 10-SNP and 592K-SNP PRS scores, respectively, included in the all-predictive factors combined model. The VIF in the Icelandic study group is somewhat smaller, compared to the other two study groups, of 1.64 and 1.63 for the 10-SNP and 592K-SNP PRS scores. The VIF indicates how much larger the variance is, compared to what it would be if the respective variables included in the model were

not correlated with each other. Moreover we estimated that the 10 SNPs under study explained ~8% of the familial risk of thyroid cancer in the Ohio study group.

Assessing thyroid cancer risk by 10-SNP PRS percentile groups based on the meta-analysis results from Ohio, Iceland and the UK. We meta-analyzed results from Ohio, Iceland and the UK after ranking 10-SNP PRS scores for each individual and correlating it with the cancer status. Individuals in the top decile group of the PRSs have a 6.9-fold risk compared to the bottom decile group ($P=5.1\times10^{-54}$; Fig 2). This difference is substantial and might be useful in clinical counseling.

Discussion

Our findings largely confirm a previous estimate that 11% of the genetic predisposition to PTC could be accounted for by the interaction of 5 common SNPs (4). In our present study involving 10 SNPs and larger numbers of cases the proportion was slightly smaller (8%); this can probably be explained by a somewhat different sample set and higher genetic resolution applied in the present study. Nevertheless, the fraction of predisposition accounted for is quite low even when ~592,000 common SNPs were investigated.

Interestingly, our data indicate that the 10-SNP PRS in the prediction model of thyroid cancer performs equally well compared to the combined model with 10-SNP PRS and common 592K-SNP PRS in all three study populations. Our observations

further support the notion that the ten variants previously detected by GWAS are important genetic factors conferring thyroid cancer risk (16). The majority of the variants (9 out of 10) are located either intronic or intergenic, while only one SNP rs6793295 is a missense variant in the *LRRC34* gene in 3q26.2 (16). This coding variant is also significantly associated with the risk of multiple myeloma, monoclonal gammopathy and interstitial lung disease (25, 26). Interestingly five non-coding variants (rs11693806, rs2466076, rs1588635, rs368187 and rs116909374) are also associated with serum levels of thyroid function related hormones (TSH, T3 and T4) and the SNP rs116909374 is associated with hypothyroidism (12, 13, 16, 27). The intergenic non-coding variant rs7902587 in 10q24.33 is significantly associated with lung cancer and ovarian cancer and rs11693806 in 2q35 is associated with breast cancer (28-30).

Our data suggest that the current PRS models with either 10-SNP or 592K-SNP PRSs could still have the problem of missing heritability (31, 32). Most likely, only few other common variants may remain to be discovered as the 592K-SNP PRS was designed to assess and estimate the contribution of such variants (16). We therefore hypothesize that hitherto undetected low frequency or rare DNA variants, in particular those located in regions of low linkage disequilibrium (LD), may play a role in PTC risk prediction (33). Indeed, we and others have demonstrated that there is a high degree of genetic heterogeneity in thyroid cancer (34-36). We have identified multiple rare or very low frequency DNA variants which may contribute to the predisposition of familial and sporadic PTC (35-38). Identification of additional low frequency and/or rare germline

DNA variants may benefit the assessment of additive genetic effects in PRS models and personalized medical diagnosis and treatment (32). We note with great interest that the recent studies by Vogelstein et al. have reached the same conclusion in that the great majority of the driver mutations in PTC are somatic events occurring randomly in stem cells of the target organ (39, 40). These findings appear to predict that only few or very few additional high-penetrance germline variants will be found in the future. Nevertheless the data presented here further support that individual PTC-associated variants confer small or modest disease risk, but the combined effect of the known associated SNPs on PTC risk can be substantial in predicting cancer risk (4, 16, 20). Our data provide evidence that PRS could be used for profiling individuals in the highest and lowest relative risk groups for thyroid cancer, which has potential for the development of population-based risk screening and stratification programs as demonstrated in other cancers (7, 41-44).

The strength of our current study is the availability of large numbers of case-control samples from three study groups representing different areas of Caucasians (Ohio, Iceland and UK). We used effect sizes obtained from meta-analyses of the Iceland/UK study groups, Ohio/UK study groups, and Ohio/Iceland study groups to estimate PRSs in the excluded population, Ohio, Iceland and UK, respectively. The PRSs constructed in this study omitted any confounding effect from the study populations being used to evaluate the correlation between the PRS and the disease status. Overall, the data presented here provide further evidence that PRS exhibits strong association with thyroid cancer. Fritsche et al reported association of PRSs in

multiple cancers, including thyroid cancer, using a PRS with 8 SNPs (7). Interestingly, they found an attenuated association between increasing thyroid cancer PRS and reduced risk for hypothyroidism (7).

In our study, thyroid cancer patients belonging to the top decile of the 10-SNP PRS have a close to 7-fold risk, relative to the bottom decile, based on meta-analysis results including data from Ohio, Iceland and the UK. We conclude that hereditary germline variants should be taken into account alongside the traditional high-penetrance variants/somatic mutations that have already become standard of care in the clinical handling of PTC (45-47). Identifying individuals with high genetic risk may prove useful to optimize screening for thyroid cancer.

Materials and Methods

Study populations. The thyroid cancer meta-analysis has been described previously (16). In short, the meta-analysis included totally 3,001 non-medullary thyroid cancer patients and 287,550 controls coming from Iceland, Ohio and Texas in the US, the Netherlands, and Spain (*SI Appendix*, Table S1). In the present study we added thyroid cancer GWAS data from the UKB (accessed under Application Number: 24711); comprised of samples from 534 patients with ICD code = C73 (PTC, FTC, Cancer/Carcinoma and rare non-medullary) and 407,945 controls, not known to have thyroid cancer (*SI Appendix,* Table S1) (22).

Genotyping. The genotyping and imputation for the study groups from Iceland, Ohio and Texas in the US, the Netherlands, and Spain has been previously described (16). Genotyping of UKB samples was performed using a custom-made Affymetrix chip, UK BiLEVE Axiom (48), and with the Affymetrix UK Biobank Axiom array (49). Imputation was performed by the Wellcome Trust Centre for Human Genetics using the Haplotype Reference Consortium (HRC) and the UK10K haplotype resources (49). This yielded a total of 96 million imputed variants, however only 40 million variants imputed using the HRC reference set were used in this study due to quality issues with the remaining variants.

Variants in the UK imputation dataset were mapped to NCBI Build38 positions and matched to the variants in the Icelandic dataset based on allele variation. The results from all study groups were combined using a fixed effect model in which the study groups were allowed to have different population frequencies for alleles and genotypes but were assumed to have a common OR and weighted with the inverse of the variance. Heterogeneity (P_{het}) was tested by comparing the null hypothesis of the effect being the same in all populations to the alternative hypothesis of each population having a different effect using a likelihood ratio test. I^2 lies between 0 and 100% and describes the proportion of total variation in study estimates that is due to heterogeneity.

Polygenic Risk Score. To evaluate the additive genetic effect of variants, we created polygenic risk scores (PRSs) as the sum of effects of each allele representing selected

sets of variants as described in Vilhjalmsson et al (24). For the PRS analysis, we regenerated three meta-analysis datasets, each time excluding data from the study group in which we intended to assess correlation between the PRSs and affection status (i.e. when generating PRSs for Icelanders we used effect estimates from a meta-analysis after excluding Icelandic results).

PRSs were generated using two different sets of variants: a) the10 published GWAS thyroid cancer risk variants, and b) using 592,475 common variants based on the previously published meta-analysis (16) including the addition of the UKB data described above. All PRS scores were standardized to have a unit standard deviation. Odds ratio per unit standard deviation increase is reported.

Statistical Analysis. Logistic regression analysis was used to assess the association of PRSs with thyroid cancer status, adjusting for year of birth, gender, ancestry with 10 principal components and familiality based on self-reported first or second-degree relative information. Familiality information was not available for the UKB samples. Strength of prediction models to predict thyroid cancer against controls was assessed by comparing the area under the curve (AUC) of the respective receiver operating curves (ROC) that plots the true-positive rate against the false-positive rate. ROC curves were compared by applying DeLong's test (50). Higher AUC indicates better model performance. We examined all pairwise correlations and calculated variance inflation factors (VIFs) associated with the PRS models (51). PRS percentile groups

were used to create categorical predictors, and risk of thyroid cancer between percentile groups was assessed applying logistic regression. Odds ratios (OR) per one SD increase to estimate the associations and AUCs to assess the discriminatory accuracy are presented with 95% CI. Familial relative risk assessment is estimated with 10-SNPs assuming an overall familial relative risk of 8.48 for thyroid cancer (16, 52, 53).

Data Availability. The authors declare that the data supporting the findings of this study are available within the article, its Supplementary, and upon reasonable request addressed to the corresponding author(s). The UK Biobank data can be obtained upon application (ukbiobank.ac.uk).

Acknowledgements. We thank Jan Lockman and Barbara Fersch for administrative help. This work was supported by National Cancer Institute Grants P30CA16058 and P01CA124570.

References

1. Xing M, Haugen BR, & Schlumberger M (2013) Progress in molecular-based management of differentiated thyroid cancer. *The Lancet* 381(9871):1058-1069.
2. Stenson PD*, et al.* (2017) The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Human Genetics* 136(6):665-677.

3. Sokolenko AP & Imyanitov EN (2018) Molecular Diagnostics in Clinical Oncology. *Front Mol Biosci* 5:76-76.
4. Liyanarachchi S*, et al.* (2013) Cumulative Risk Impact of Five Genetic Variants Associated with Papillary Thyroid Carcinoma. *Thyroid* 23:1532-1540.
5. Szulkin R*, et al.* (2015) Prediction of individual genetic risk to prostate cancer using a polygenic score. *The Prostate* 75(13):1467-1474.
6. Maas P*, et al.* (2016) Breast Cancer Risk From Modifiable and Nonmodifiable Risk Factors Among White Women in the United StatesModifiable and Nonmodifiable Risk Factors and Breast Cancer RiskModifiable and Nonmodifiable Risk Factors and Breast Cancer Risk. *JAMA Oncology* 2(10):1295-1302.
7. Fritsche LG*, et al.* (2018) Association of Polygenic Risk Scores for Multiple Cancers in a Phenome-wide Study: Results from The Michigan Genomics Initiative. *The American Journal of Human Genetics* 102(6):1048-1061.
8. Mavaddat N*, et al.* (2019) Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. *The American Journal of Human Genetics* 104(1):21-34.
9. Goldgar DE, Easton DF, Cannon-Albright LA, & Skolnick MH (1994) Systematic population-based assessment of cancer risk in first-degree relatives of cancer probands. *J Natl Cancer Inst* 86(21):1600-1608.
10. Dong C & Hemminki K (2001) Modification of cancer risks in offspring by sibling and parental cancers from 2,112,616 nuclear families. *Int J Cancer* 92(1):144-150.

11. Risch N (2001) The genetic epidemiology of cancer: interpreting family and twin studies and their implications for molecular genetic approaches. *Cancer Epidemiol Biomarkers Prev* 10(7):733-741.
12. Gudmundsson J*, et al.* (2009) Common variants on 9q22.33 and 14q13.3 predispose to thyroid cancer in European populations. *Nat Genet* 41(4):460–464.
13. Gudmundsson J*, et al.* (2012) Discovery of common variants associated with low TSH levels and thyroid cancer risk. *Nat Genet* 44(3):319-322.
14. Köhler A*, et al.* (2013) Genome-Wide Association Study on Differentiated Thyroid Cancer. *The Journal of Clinical Endocrinology & Metabolism* 98(10):E1674-E1681.
15. Son H-Y*, et al.* (2017) Genome-wide association and expression quantitative trait loci studies identify multiple susceptibility loci for thyroid cancer. *Nat Commun* 8:15966.
16. Gudmundsson J*, et al.* (2017) A genome-wide association study yields five novel thyroid cancer risk loci. *Nat Commun* 8:14517.
17. Figlioli G*, et al.* (2014) Novel Genome-Wide Association Study–Based Candidate Loci for Differentiated Thyroid Cancer Risk. *The Journal of Clinical Endocrinology & Metabolism* 99(10):E2084-E2092.
18. Figlioli G*, et al.* (2016) A Comprehensive Meta-analysis of Case–Control Association Studies to Evaluate Polymorphisms Associated with the Risk of Differentiated Thyroid Carcinoma. *Cancer Epidemiology Biomarkers & Prevention* 25(4):700.

19. Mancikova V*, et al.* (2015) Thyroid cancer GWAS identifies 10q26.12 and 6q14.1 as novel susceptibility loci and reveals genetic heterogeneity among populations. *International Journal of Cancer* 137(8):1870-1878.

20. Figlioli G*, et al.* (2015) Novel genetic variants in differentiated thyroid cancer and assessment of the cumulative risk. *Scientific Reports* 5:8922.

21. Hwangbo Y*, et al.* (2018) Genome-Wide Association Study Reveals Distinct Genetic Susceptibility of Thyroid Nodules From Thyroid Cancer. *The Journal of Clinical Endocrinology & Metabolism* 103(12):4384-4394.

22. Bycroft C*, et al.* (2017) Genome-wide genetic data on ~500,000 UK Biobank participants. *bioRxiv*:166298.

23. Bycroft C*, et al.* (2018) The UK Biobank resource with deep phenotyping and genomic data. *Nature* 562(7726):203-209.

24. Vilhjálmsson Bjarni J*, et al.* (2015) Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *The American Journal of Human Genetics* 97(4):576-592.

25. Swaminathan B*, et al.* (2015) Variants in ELL2 influencing immunoglobulin levels associate with multiple myeloma. *Nature Communications* 6(1):7213.

26. Fingerlin TE*, et al.* (2013) Genome-wide association study identifies multiple susceptibility loci for pulmonary fibrosis. *Nature Genetics* 45(6):613-620.

27. Kichaev G*, et al.* (2019) Leveraging Polygenic Functional Enrichment to Improve GWAS Power. *The American Journal of Human Genetics* 104(1):65-75.

28. McKay JD*, et al.* (2017) Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. *Nature Genetics* 49(7):1126-1132.

29. Phelan CM*, et al.* (2017) Identification of 12 new susceptibility loci for different histotypes of epithelial ovarian cancer. *Nature Genetics* 49(5):680-691.

30. Michailidou K*, et al.* (2017) Association analysis identifies 65 new breast cancer risk loci. *Nature* 551(7678):92-94.

31. Zuk O, Hechter E, Sunyaev SR, & Lander ES (2012) The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences* 109(4):1193.

32. Young AI (2019) Solving the missing heritability problem. *PLOS Genetics* 15(6):e1008222.

33. Wainschtein P*, et al.* (2019) Recovery of trait heritability from whole genome sequence data. *bioRxiv*:588020.

34. Lesueur F*, et al.* (1999) Genetic heterogeneity in familial nonmedullary thyroid carcinoma: exclusion of linkage to RET, MNG1, and TCO in 56 families. NMTC Consortium. *J Clin Endocrinol Metab* 84(6):2157-2162.

35. He H*, et al.* (2013) SRGAP1 Is a Candidate Gene for Papillary Thyroid Carcinoma Susceptibility. *Journal of Clinical Endocrinology & Metabolism* 98(5):E973-E980.

36. He H*, et al.* (2013) Ultra-rare mutation in long-range enhancer predisposes to thyroid carcinoma with high penetrance. *PLoS One* 8(5):e61920.

37. Tomsic J*, et al.* (2015) A germline mutation in SRRM2, a splicing factor gene, is implicated in papillary thyroid carcinoma predisposition. *Scientific Reports* 5:10566.

38. Wang Y*, et al.* (2019) Identification of Rare Variants Predisposing to Thyroid Cancer. *Thyroid* 29(7):946-955.

39. Tomasetti C & Vogelstein B (2015) Cancer etiology. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science (New York, N.Y.)* 347(6217):78-81.

40. Tomasetti C, Li L, & Vogelstein B (2017) Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science* 355(6331):1330.

41. Mavaddat N*, et al.* (2015) Prediction of breast cancer risk based on profiling with common genetic variants. *Journal of the National Cancer Institute* 107(5):djv036.

42. Frampton M & Houlston RS (2017) Modeling the prevention of colorectal cancer from the combined impact of host and behavioral risk factors. *Genetics in medicine : official journal of the American College of Medical Genetics* 19(3):314-321.

43. Radice P, Pharoah PDP, & Peterlongo P (2016) Personalized testing based on polygenic risk score is promising for more efficient population-based screening programs for common oncological diseases. *Annals of Oncology* 27(3):369-370.

44. Yang X*, et al.* (2018) Evaluation of polygenic risk scores for ovarian cancer risk prediction in a prospective cohort study. *Journal of Medical Genetics* 55(8):546.

45. Nikiforov YE (2017) ROLE OF MOLECULAR MARKERS IN THYROID NODULE MANAGEMENT: THEN AND NOW. *Endocrine Practice* 23(8):979-988.

46. Nikiforova MN*, et al.* (2018) Analytical performance of the ThyroSeq v3 genomic classifier for cancer diagnosis in thyroid nodules. *Cancer* 124(8):1682-1690.

47. Endo M*, et al.* (2019) Afirma Gene Sequencing Classifier Compared with Gene Expression Classifier in Indeterminate Thyroid Nodules. *Thyroid* 29(8):1115-1124.

48. Wain LV*, et al.* (2015) Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank. *The Lancet Respiratory Medicine* 3(10):769-781.

49. Welsh S, Peakman T, Sheard S, & Almond R (2017) Comparison of DNA quantification methodology used in the DNA extraction protocol for the UK Biobank cohort. *BMC Genomics* 18(1):26.

50. DeLong ER, DeLong DM, & Clarke-Pearson DL (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 44(3):837-845.

51. Fox J & Monette G (1992) Generalized Collinearity Diagnostics. *Journal of the American Statistical Association* 87(417):178-183.

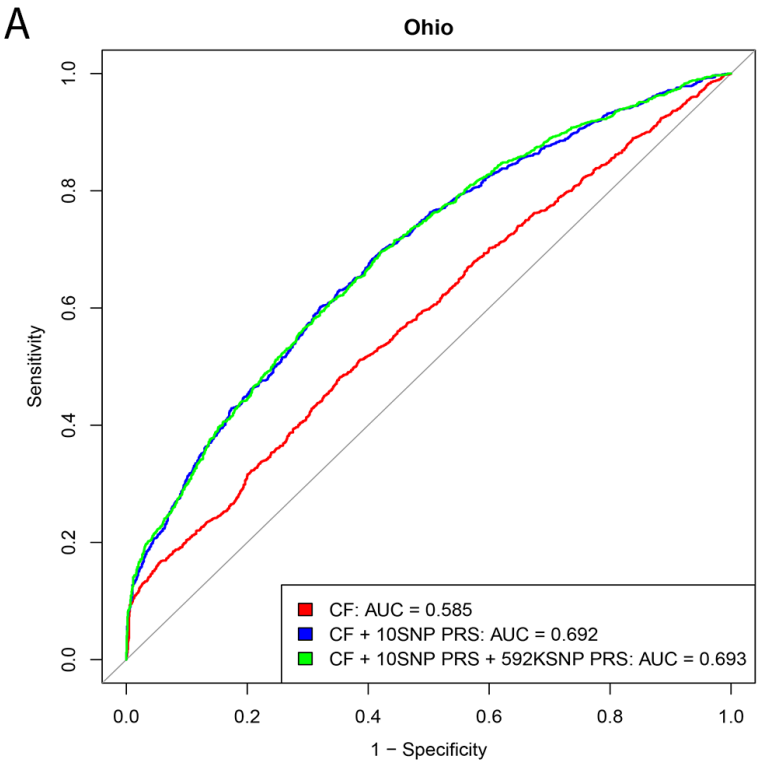
52. Houlston RS & Ford D (1996) Genetics of coeliac disease. *QJM : monthly journal of the Association of Physicians* 89(10):737-743.

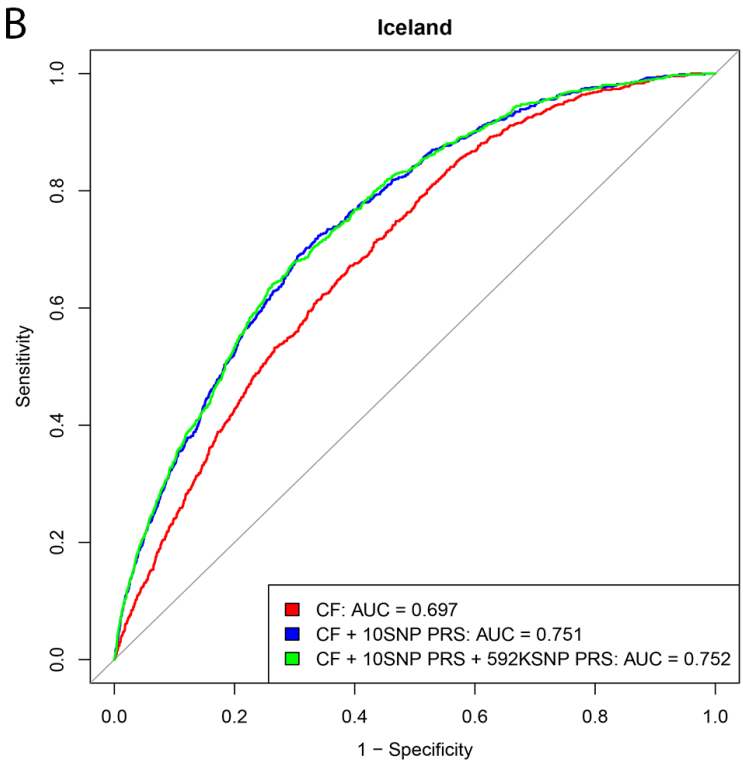
53. Broderick P*, et al.* (2011) Common variation at 3p22.1 and 7p15.3 influences multiple myeloma risk. *Nature Genetics* 44:58.

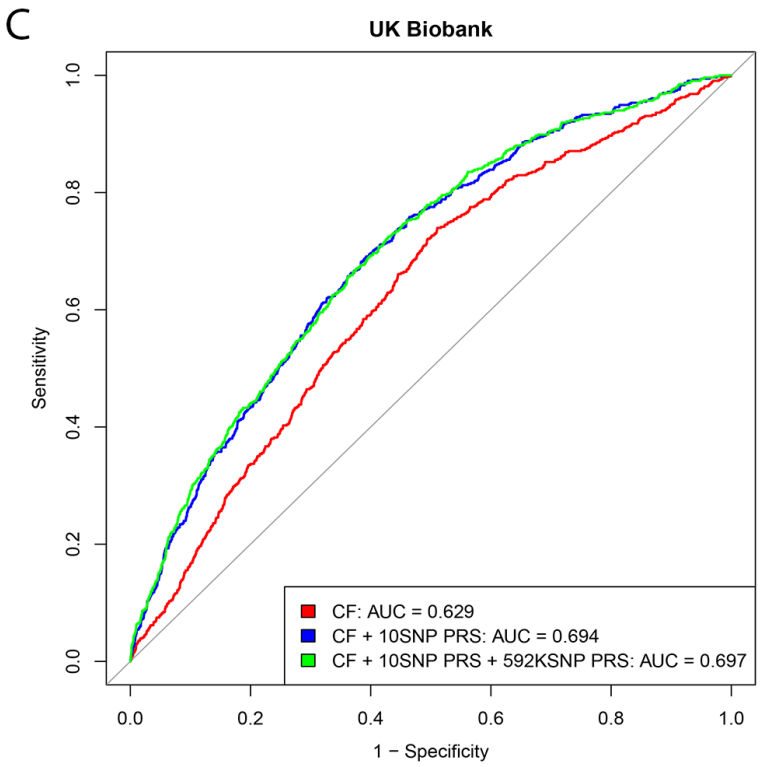
Figure Legends

Figure 1. Receiver–operator characteristic (ROC) curves assessing the discriminative power of the PRS models. (*A*) The Ohio study group. (*B*) The Iceland study group. (*C*) The UK study group. CF, Clinical factors model with year of birth, gender, ancestry, and familiality except for the UK where no information was available about family history of thyroid cancer.

Figure 2. Odds ratio (OR) estimates for 10-SNP PRS deciles of thyroid cancer status obtained from the meta-analysis results from Ohio, Icelandic, and the UK study groups; using the bottom 10-SNP PRS decile (0%–10%) as the reference group (shown as a horizontal solid line).


A
Ohio
Sensitivity
1 − Specificity
0.0
0.2
0.4
0.6
0.8
1.0
CF: AUC = 0.585
CF + 10SNP PRS: AUC = 0.692
CF + 10SNP PRS + 592KSNP PRS: AUC = 0.693


B
Iceland
Sensitivity
1 − Specificity
0.0
0.2
0.4
0.6
0.8
1.0
CF: AUC = 0.697
CF + 10SNP PRS: AUC = 0.751
CF + 10SNP PRS + 592KSNP PRS: AUC = 0.752


C
UK Biobank
Sensitivity
1 − Specificity
0.0
0.2
0.4
0.6
0.8
1.0
CF: AUC = 0.629
CF + 10SNP PRS: AUC = 0.694
CF + 10SNP PRS + 592KSNP PRS: AUC = 0.697

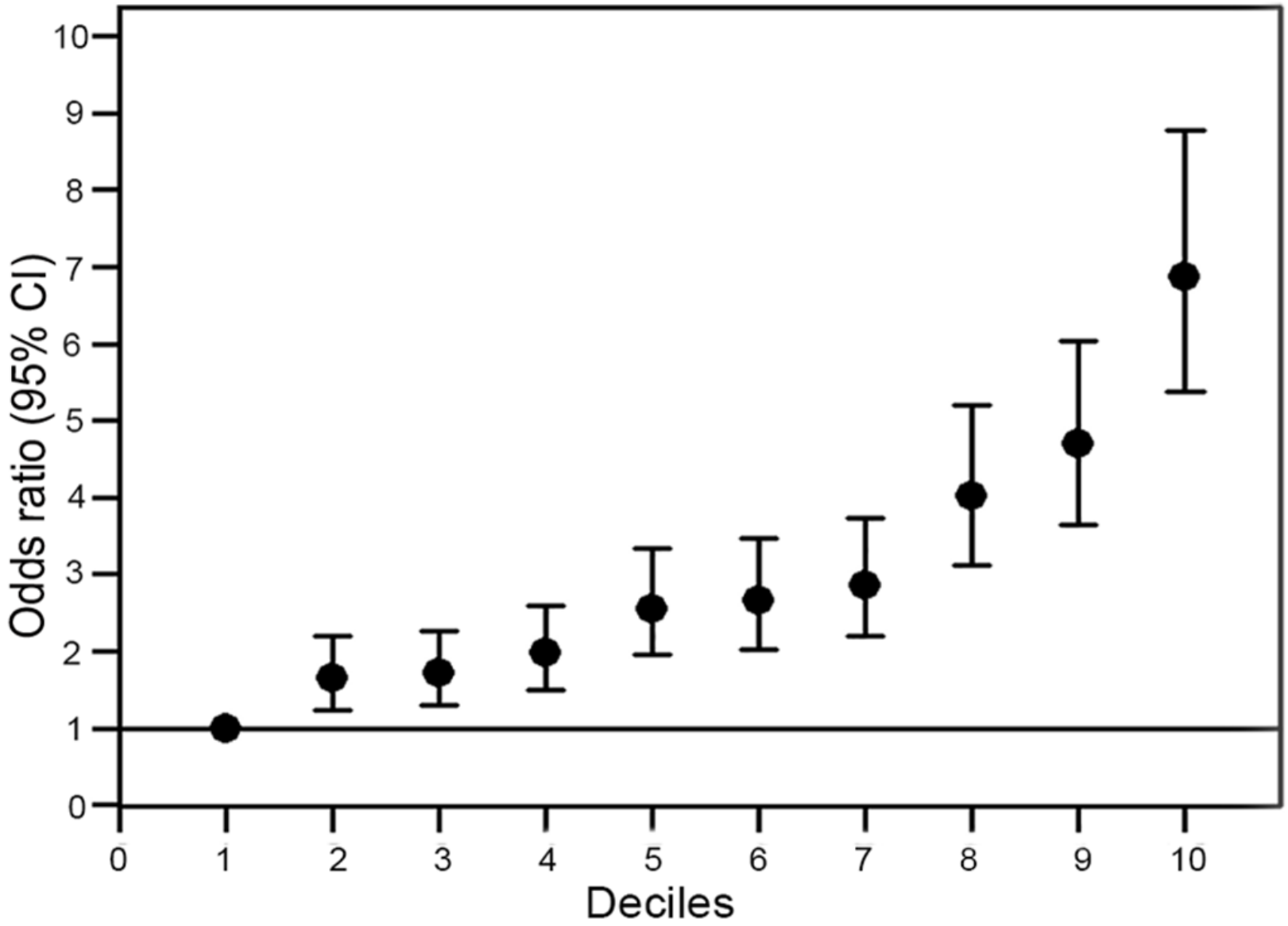

Odds ratio (95% CI)
10
9
8
7
6
5
4
3
2
1
0
0
1
2
3
4
5
6
7
8
9
10
Deciles

Table 1. Summary of demographic characteristics of the Ohio, Iceland and UKB study groups

Study group	Characteristic	Patients	Controls
Ohio	Total	1,544	1,593
	Gender		
	Male n(%)	395 (26)	423 (27)
	Female n(%)	1,149 (74)	1,170 (73)
	Mean age (sd.)	42.9 (±15.1)	45.2 (±14.0)
	Median year of birth (range)	1962 (1913-2001)	1963 (1918-1991)
	First or Second degree relative diagnosed with thyroid cancer		
	Yes n(%)	135 (8.7)	9 (0.6)
	No n(%)	1,409 (91.3)	1,584 (99.4)
Iceland	Total	723	129,556
	Gender		
	Male n(%)	183 (25.3)	60,282 (46.5)
	Female n(%)	540 (74.7)	69,274 (53.5)
	Mean age (sd.)	49.3 (±17.3)	60.4 (±18.0)
	Median year of birth (range)	1946 (1911-1989)	1956 (1890-1990)
	First or Second degree relative diagnosed with thyroid cancer		
	Yes n(%)	84 (11.6)	6,455 (5.0)
	No n(%)	639 (88.4)	123,101 (95.0)
UKB	Total	534	407,945
	Gender		
	Male n(%)	131 (24.5)	187,661 (46.0)
	Female n(%)	403 (75.5)	220,818 (54.0)
	Mean age (sd.)	51.8 (±12.18)	64.1 (±8.00)
	Median year of birth (range)	1949 (1938-1969)	1950 (1934-1969)
	*First or Second degree relative diagnosed with thyroid cancer		
	Yes n(%)	na	na
	No n(%)	na	na

*No information about family history of thyroid cancer is available for the UKB samples

Table 2. Effect estimates used in PRS model in each study group.

					Ohio		Iceland		UKB	
Marker	Locus	Position (bp)*	OA	EA	EAF	OR[a]	EAF	OR[b]	EAF	OR[c]
rs12129938	1q42.2	233,276,815	G	A	0.81	1.2	0.783	1.16	0.775	1.32
rs11693806	2q35	217,427,435	G	C	0.318	1.43	0.285	1.37	0.279	1.43
rs6793295	3q26.2	169,800,667	C	T	0.755	1.2	0.78	1.16	0.729	1.23
rs73227498	5q22.1	112,150,207	T	A	0.891	1.28	0.855	1.33	0.863	1.37
rs2466076	8p12	32,575,278	T	G	0.528	1.32	0.467	1.3	0.452	1.32
rs1588635	9q22.33	97,775,520	C	A	0.476	1.64	0.356	1.72	0.326	1.69
rs7902587	10q24.33	103,934,543	C	T	0.12	1.25	0.095	1.22	0.098	1.41
rs368187	14q13.3	36,063,370	C	G	0.626	1.33	0.523	1.28	0.55	1.39
rs116909374	14q13.3	36,269,155	C	T	0.044	1.71	0.049	1.73	0.038	1.71
rs2289261	15q22.33	67,165,147	G	C	0.7	1.14	0.669	1.22	0.647	1.23

* the chromosomal position in base pairs (bp) with reference to Build 38, the effect allele (EA) and the other allele (OA), the effect allele frequency (EAF) in controls for each study group, and the odds ratio (OR) used to calculate the polygenic risk score for each study group.

[a]Odds ratio from the meta analysis after excluding results for the Ohio study group

[b]Odds ratio from the meta analysis after excluding results for the Icelandic study group

[c]Odds ratio from the meta analysis after excluding results for the UKB study group

Table 3. Classification results for different models in each study group

Study group	Model	AUC	95% CI	P-value*
Ohio	CF	0.585	(0.565-0.605)	Reference
	CF+10SNPs_PRS	0.692	(0.673-0.710)	3.10E-21
	CF+10SNPs_PRS+592KSNPs_PRS	0.693	(0.675-0.712)	0.34
Iceland	CF	0.697	(0.680-0.714)	Reference
	CF+10SNPs_PRS	0.751	(0.736-0.768)	3.00E-14
	CF+10SNPs_PRS+592KSNPs_PRS	0.752	(0.680-0.714)	0.3
UKB	CF	0.629	(0.606-0.651)	Reference
	CF+10SNPs_PRS	0.694	(0.673-0.716)	1.00E-09
	CF+10SNPs_PRS+592KSNPs_PRS	0.697	(0.676-0.719)	0.27

*The P-value is for the stepwise addition of factors shown in the Model column