Genetic signature of differentiated thyroid carcinoma susceptibility: a machine learning approach

 Accepted Manuscript published as ETJ-22-0058.R2. Accepted for publication: 17-Aug-2022

Genetic signature of differentiated thyroid carcinoma susceptibility: a machine learning approach

Giulia Brigante[1,2], Clara Lazzaretti[1], Elia Paradiso[1], Federico Nuzzo[1], Martina Sitti[1], Frank Tüttelmann[3], Gabriele Moretti[4], Roberto Silvestri[4], Federica Gemignani[4], Asta Försti[5,6], Kari Hemminki[7,8], Rossella Elisei[9], Cristina Romei[9], Eric Adriano Zizzi[10], Marco Agostino Deriu[10], Manuela Simoni[1,2,11], Stefano Landi[4,*], Livio Casarini[1,11,*]

1. Unit of Endocrinology, Department of Biomedical, Metabolic and Neural Sciences, University of Modena and Reggio Emilia, Modena, Italy

2. Unit of Endocrinology, Department of Medical Specialties, Azienda Ospedaliero-Universitaria, Modena, Italy

3. Institute of Reproductive Genetics, University of Münster, Münster, Germany

4. Department of Biology, University of Pisa, Pisa, Italy

5. Hopp Children's Cancer Center (KiTZ), Heidelberg, Germany

6. Division of Pediatric Neurooncology, German Cancer Research Center (DKFZ), German Cancer Consortium (DKTK), Heidelberg, Germany

7. Biomedical Center, Faculty of Medicine and Biomedical Center in Pilsen, Charles University in Prague, 30605 Pilsen, Czech Republic

8. Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 580, D-69120, Heidelberg, Germany

9. Department of Endocrinology, University Hospital, Pisa, Italy

10. Polito^{BIO}Med Lab, Department of Mechanical and Aerospace Engineering, Politecnico di Torino, Italy

11. Center for Genomic Research, University of Modena and Reggio Emilia, Modena, Italy

*Corresponding authors: Professor Stefano Landi. Department of Biology, University of Pisa, Pisa, Italy. Email: stefano.landi@unipi.it; Professor Livio Casarini. Unit of Endocrinology, Department of Biomedical, Metabolic and Neural Sciences, University of Modena and Reggio Emilia, Modena, Italy. Email: livio.casarini@unimore.it

ABSTRACT

To identify a peculiar genetic combination predisposing to differentiated thyroid carcinoma (DTC), we selected a set of single-nucleotide polymorphisms (SNPs) associated with DTC risk, considering polygenic risk score (PRS), Bayesian statistics, and a machine learning (ML) classifier to describe cases and controls in 3 different datasets.

Dataset 1 (649 DTC, 431 controls) has been previously genotyped in a genome-wide association study (GWAS) on Italian DTC. Dataset 2 (234 DTC, 101 controls) and dataset 3 (404 DTC, 392 controls) were genotyped. Associations of 171 SNPs reported to predispose to DTC in candidate studies were extracted from the GWAS of dataset 1, followed by replication of SNPs associated with DTC risk ($P<0.05$) in dataset 2. The reliability of the identified SNPs was confirmed by PRS and Bayesian statistics after merging the three datasets. SNPs were used to describe the case/control state of individuals by ML classifier.

Starting from 171 SNPs associated with DTC, 15 were positive in both the datasets 1 and 2. Using these markers, PRS revealed that individuals in the fifth quintile had a 7-fold increased risk of DTC than those in the first. Bayesian inference confirmed that the selected 15 SNPs differentiate cases from controls. Results were corroborated by ML, finding a maximum AUC of about 0.7.

A restricted selection of only 15 DTC-associated SNPs is able to describe the inner genetic structure of Italian individuals and ML allows a fair prediction of case or control status based solely on the individual genetic background.

 Accepted Manuscript published as ETJ-22-0058.R2. Accepted for publication: 17-Aug-2022

INTRODUCTION

Thyroid cancer is the most common endocrine neoplasia with a worldwide estimated age-standardized incidence rate of 6.7 per 100000 in 2018 [1]. Differentiated thyroid carcinoma (DTC) is the most frequent sub-type of thyroid cancer with increasing incidence in the last 20 years, likely because of the increased knowledge of associated risk factors and ameliorated diagnostic procedures [2]. However, most DTC have a favourable prognosis [3] and the diagnostic-therapeutic procedures should aim to avoid both delayed diagnosis and overmedication.

To date, the management of thyroid nodules suspected to be DTC is mainly guided by the sonographic risk pattern and the coexistence of other risk factors [3]. Genetics could play a role in helping the diagnostic process, assuming the possibility to stratify patients according to a personalized risk profile [4]. This stems from the observation that blood relatives of patients diagnosed with DTC show highly increased risk for the disease, implying the existence of an important genetic component [5,6]. The role of genes in the aetiology of DTC has been studied in populations and most of the risk alleles have been identified by case-control and genome-wide association studies (GWAS) [7-15]. However, it is still difficult to predict the individual risk of DTC based on the existing data, likely because of a complex interaction among multiple co-inherited low/moderate penetrant alleles. In fact, one single common variant *per se* is weakly associated with increased DTC risk, which could instead emerge as a cumulative effect of several single nucleotide polymorphisms (SNPs) with individual low impact. Thus, the overall risk could be the result of complex gene-gene and gene-environment interactions.

In order to take into account multiple alleles, the measure of disease susceptibility could be provided by calculating the polygenic risk score (PRS), where each variant allele is treated as an individual, independent, risk factor and subjects are stratified according to the number of risk alleles, in additive or weighted models. The so calculated cancer risk may achieve relatively high odd ratio (OR) values [16] [17-22]. For DTC, it has been shown that people carrying ≥14 risk alleles have an about 8-fold increased risk compared to people carrying ≤7 risk alleles [23,24]. Therefore, the PRS is a promising

method for risk prediction. However, gene-gene interactions are likely too complex to be explained by simple additive or weighted models and alternative methods are under exploration.

Machine learning (ML) is increasingly used for predicting individuals' inherited genomic susceptibility to cancer [25]. Another interesting approach is represented by Bayesian statistics for population genetics, in which individuals are assigned to ethnic sub-groups or phenotypes according to their underlying genetic structure [26,27]. Genetic data may serve to run ML diagnostic analyses aimed at stratifying individuals into disease risk categories [28]. However, these methods have not been fully exploited for dissecting complex traits, such as the susceptibility to cancer, assuming it as a phenotype information. To the best of our knowledge, ML has never been applied before to the study of genetic predisposition to DTC.

In this replication study, we aimed to assess the genetic signatures associated with the predisposition to DTC. For this purpose, a small number of single nucleotide polymorphisms (SNPs) descriptive of a DTC-related genotype were selected in three independent genetic datasets and confirmed by Bayesian statistics. The diagnostic performance of the selected markers in categorizing the case/control state of subjects was evaluated by ML techniques.

 Accepted Manuscript published as ETJ-22-0058.R2. Accepted for publication: 17-Aug-2022

METHODS

Study design

Briefly, we identified a relatively low number of SNPs highly associated with DTC by sequential association analyses in three independent case/control series. These SNPs served as a genetic information to describe the case/control status of Italian individuals by ML methods. Firstly, a selection of 171 candidate DTC-associated SNPs was obtained from the literature (see paragraph "SNPs selection"). The SNP list was further reduced after testing for SNPs association with the disease. To this purpose, genetic data from two of the available datasets (dataset 1 and 2; Figure 1a), each comprising Italian DTC subjects and healthy controls, were used. Briefly, 34 SNPs considered significantly associated with DTC in dataset 1 were genotyped *ad hoc* and checked for relevance in the independent dataset 2. SNP selection criteria are reported in detail in the paragraph "Statistical analysis". Finally, a total of 15 SNPs highly associated with DTC in both datasets were obtained and further genotyped *ad hoc* in the independent dataset 3. Their potential of describing DTC signature was confirmed by a control Bayesian clustering in the merged three datasets (see paragraph "Bayesian statistics for population genetics"). ML methods were run to confirm the case/control state of individuals, using the selected 15 SNPs as input variables. To this purpose, an extended dataset was built by merging the two largest datasets (1 and 3), to obtain a pool of randomly chosen "training" (80% of the merged dataset) and "testing data" (20%). After finding the most effective ML algorithms, a replication analysis was set on the dataset 2. The whole procedure is summarized in Figure 1b.

[FIGURE 1]

Subjects

Dataset 1 has been previously described in a GWAS on DTC [12]. It included Italian DTC cases and controls recruited consecutively from the Department of Endocrinology, University Hospital of Pisa, Italy in the period January 2009-August 2011 [12]. Overall, the genotypes of 649 DTC patients and 431 healthy controls were considered. Dataset 2 included 234 Italian DTC patients and 101 healthy

controls recruited at the Unit of Endocrinology, University Hospital of Modena, Italy, between 2008 and 2012. These individuals were genotyped for 34 DTC-associated SNPs ("SNP selection" section) after DNA extraction from blood samples (Supplementary text). Dataset 3 included 404 DTC subjects and 392 controls recruited at the Department of Endocrinology, University Hospital of Pisa, Italy, between September 2011 and December 2012, and subjected to genotyping for 15 DTC-associated SNPs ("SNP selection" section) after extraction of DNA from blood. All the DTC diagnoses have been histologically confirmed after thyroidectomy. Controls were recruited among healthy volunteers without known thyroid disease and/or with negative thyroid ultrasound. In details, controls of Dataset 1 and 3 comprised healthy individuals without known thyroid disease recruited during a routine health screening, or blood donor volunteers. Controls of Dataset 2 were volunteers recruited by local advertisement as the control group for an ongoing case-control study on thyroid cancer; one of the participants had a personal history of thyroid disease and they had never undergone any thyroid ultrasound scan before; they performed thyroid ultrasound and thyroid resulted to be normal for size, position and echogenicity, without cystic or nodular lesions.

All the subjects enrolled in the three independent datasets were unrelated.

Information about sex, age at diagnosis of DTC for cases and age at recruitment for controls, anthropometric measurements (height and weight) were collected. Body-mass index (BMI) was also calculated as the weight (kg)/height $(m)^2$ ratio. Individuals underwent peripheral blood withdrawn and samples were stored at -20°C until analysis. DNA was extracted from EDTA-venous blood samples using standard methodologies. In dataset 1, SNPs missing in the GWAS were obtained by imputation by exploiting the linkage disequilibrium (LD) blocks [29]. SNP genotyping in datasets 2 and 3 was performed with the iPLEX® assay (Life & Brain GmbH, Bonn, Germany) (Supplementary text).

The local Ethics Committees of Modena and Pisa (Italy) approved the study (Protocol Nr. 122/08, Nr. 7116/09 and Nr. 2359/14) and all participants signed a written informed consent.

SNP selection

Accepted Manuscript published as ETJ-22-0058.R2. Accepted for publication: 17-Aug-2022

We considered all the SNPs associated with DTC on the PubMed database using the following keywords alone and/or in different combinations: papillary thyroid cancer, thyroid cancer, thyroid tumour, DTC, PTC, GWAS, association. A total of 171 SNPs were initially selected from 156 studies, including both candidate gene studies and GWAS, demonstrating an association with DTC ($P<0.05$) (Supplementary table S1). These SNPs were evaluated for their association with DTC risk in the dataset 1 (Supplementary text; Supplementary table S2). A subset of 34 selected SNPs successfully passed the test and they were genotyped in the dataset 2. 15 SNPs were considered positive (Supplementary table S3) and genotyped in the dataset 3. These SNPs were used for the Bayesian analysis of population genetic structure and assessment of genetics disease risk using ML algorithms. The selection criteria are shown (Figure 2) and further explained in the "Statistical analysis" section.

[FIGURE 2]

Statistical analysis

Each genotype was evaluated by the chi-square test for the Hardy-Weinberg equilibrium (HWE) in controls, employing the Bonferroni's correction (P threshold=$1.47x10^{-3}$). The association between the health state and genotypes was evaluated with a multivariate logistic regression analysis (MLRA). The model returns the odds ratio adjusted (ORadj) for covariates (e.g. sex and age) and their 95% confidence intervals (CI) with a statistical P-value of the association. The most likely mode of inheritance was evaluated performing an extended MAX test [30] based on multiplicity-adjusted P-values for the Cochran–Armitage trend test of the dominant, additive and recessive models.

In order to select SNPs robustly associated with the DTC risk, among the 171 candidates, we carried out a two-stage case-control association study. The first step was performed by evaluating the extent of association of the candidate SNPs with DTC risk obtained in dataset 1 [12]. For each SNP, the additive, recessive and dominant model of inheritance were evaluated and SNPs showing a statistically significant association ($p<0.05$) were passed to the second step, performed on dataset 2. The selected SNPs served for PRS and weighted PRS (wPRS) calculation, in the three merged datasets. The PRS was built by summing the total number of risk alleles for each subject (attributing the value of 1 to each risk allele). The wPRS was built by assigning to each genotype the relative OR obtained in the GWAS. Then, the ORs were multiplied. For PRS, we assessed the cumulative effect of the independent significant SNPs with an additive model. For each SNP the genotypes were coded as 0, 1 or 2, indicating the number of risk alleles in the genotype. Then, individuals were grouped according to the total number of risk alleles into quintiles with the lowest group used as the reference. For wPRS, as previously reported [31], the number of risk alleles for each genotype was multiplied for its relative weight, based on the association of the allele with the health state, as: $PRS=\beta1\ x1+ \beta2\ x2 + [\cdots] + \beta k\ xk +\cdots+ \beta n\ xn$; where βk is the per-allele log OR for the disease associated with SNP k, xk is the allele dosage for SNP k, and n is the total number of SNPs included in the PRS.

Bayesian statistics for population genetics

 Accepted Manuscript published as ETJ-22-0058.R2. Accepted for publication: 17-Aug-2022

We tested the capability of the 15 selected SNPs to provide a DTC genetic signature in the merged datasets 1, 2 and 3. The genetic structure of DTC patients and healthy controls was explored according to methods of Bayesian statistics for population genetics implemented in the STRUCTURE 2.3.4 software [27], as previously described [32]. The case/control state of individuals was unknown to the software, which inferred genetic structures using only SNP data. Bayesian analysis and software settings are detailed in the supplementary online material (Supplementary text).

Machine Learning-based analysis

In the preliminary phase of ML algorithm selection, different approaches were tested, namely k-Nearest Neighbors (kNN), Naïve Bayes (NB) [33], Random Forest (RF), Gradient Boosting (GB) [34], AdaBoost (AB) [35] and Support Vector Machine (SVM) algorithms, as implemented in the SciKit-Learn [36] library for Python. The AdaBoost classifier [37] was selected as the best overall algorithm (Supplementary text; Supplementary figure S1). We used the SciKit-Learn implementation of the AdaBoost classifier, where the base learner is a Decision Tree classifier with a maximum depth of 1, sometimes referred to as “decision stump”. The total number of base estimators was tuned in the range 1-100 (with a step of 1), to maximize ROC-AUC on the test set. The classifier was run on three datasets (Table 1): (1) a training set, used for the training of the algorithm, composed of a randomly extracted 80% of the merged dataset 1+3; (2) a test set, for an initial performance evaluation and hyperparameter tuning, composed of the remaining 20% of the merged 1+3 dataset; (3) a validation set, corresponding to dataset 2 after pruning missing values, which constitutes a third, unseen dataset used for external validation.

[TABLE 1]

RESULTS

Population characteristics

Characteristics of subjects enrolled in the study are summarized (Table 2).

[TABLE 2]

SNPs associated with DTC

The overall workflow for identification of SNPs associated with the risk of DTC (Figure 2) is extensively descripted and results provided as an online supplementary material (Supplementary text). We selected 171 SNPs associated with DTC with a $P < 0.05$ from the online literature database (Supplementary table S1) and SNPs associated with the risk of DTC in dataset 1 with a $P < 0.05$ were considered as positive (Supplementary table S2), then were genotyped in dataset 2. All SNPs were in HWE in controls. Among these SNPs, four were robustly associated with the risk of DTC (rs965513, rs3758249, rs7048394, rs944289), as they accomplished the Bonferroni's threshold of statistical significance in the combined dataset 1 and 2 (Supplementary table S3). Three SNPs (rs6759952, rs966423, and rs1203952) were considered highly likely DTC risk markers as they were positive in both datasets at the nominal P-value of 0.05. Eight SNPs (rs10238549, rs7800391, rs1799814, rs7617304, rs4808708, rs10781500, rs1061758, and rs10877887) were considered as possible DTC risk markers, as they were statistically significant at the level of 0.05 in the combined datasets. Thus, we finally selected 15 SNPs strongly associated with DTC in datasets 1 and 2 (Table 3). None of them was in linkage disequilibrium with each other ($r^2<0.8$).

[TABLE 3]

These 15 SNPs were then genotyped in dataset 3, while the remaining 156 SNPs were considered not associated with the risk of DTCs in our study populations.

Calculation of polygenic risk scores

The 15 positive SNPs were used for the calculation of PRS and wPRS in the merged data of all three datasets. Subjects were divided in quintiles based on the number of risk alleles and the lowest quintile was used as the reference. The risk increased progressively with the increasing number of risk alleles,

 Accepted Manuscript published as ETJ-22-0058.R2. Accepted for publication: 17-Aug-2022

up to the value of OR_{adj} = 6.87 (95% CI = 4.9-9.64) for the fifth quintile in the wPRS. All the differences were highly statistically significant both in the PRS and the wPRS, already from the second and third quintile (Table 4; Supplementary figure S2).

[TABLE 4]

Exploration of individual's genetic structures according to DTC-related SNPs

The association between risk of DTC and individual genetic profile was explored by the application of Bayesian inference. The 15 SNPs used for the PRS were also employed as an input for the STRUCTURE software, run on the merged datasets 1, 2 and 3. STRUCTURE returned the relative weight of each component in the genetic background of each subject shown as a bar plot (Figure 3). Five (k=5) possible genetic structures (components) were found as the most representative of the datasets [26] (Supplementary text). The pattern, calculated using 15 SNPs, reflects at a glance the different DTC-related genetic profile between cases (DTC) and controls.

[FIGURE 3]

Output data representing the DTC-related genetic background were analysed to evaluate the quality of Bayesian inference by multiple regression analysis. We identified two components strongly associated with the case/control state: the component 2 (k2) had a F-ratio of 327.66 and the component 3 (k3) had a F-ratio of 106.26 (both $p<10^{-6}$). Results were confirmed by multivariate logistic regression analysis using the two components as continuous variables. In this case, we found that they were strongly associated with DTC risk, with ORs of 143.4 (95% CI= 52.7- 390.2) and 12.2 (95% CI=5.72-26.1).

ML-based DTC description using SNP information

The AdaBoost algorithm was found to be the most effective and well-calibrated in classifying individuals (Supplementary text; Supplementary figure S3). The classifier was further tuned in terms of the number of base estimators hyperparameter, in a range of 1 to 100. We found 25 to be an optimal number of base estimators, providing an optimal balance between computational cost and model accuracy (Supplementary text; Supplementary figure S4 and S5). Additionally, predicted probability calibration was implemented using Platt's method [38]. The detailed metrics of the AdaBoost classifier are reported (Table 5), as well as ROC curves and AUC of all datasets (Figure 4A).

[TABLE 5]

Results clearly highlight that there is no significant overfitting on the training set (Table 5; Figure 4A), given the reduced differences between the training and test set performance in terms of AUC (0.04), accuracy (0.6%), sensitivity (0.01) and specificity (0.02). This is also confirmed by the 10-fold cross-validation on the train/test splits, which resulted in an average ROC AUC of 0.65 ± 0.03 (SD). In addition, when classifying the samples from the external validation set, which again showed comparable classification performance, the model's ability to generalize on unseen data noticeably emerges. Analysis of the predicted probabilities revealed that they fairly match the real distribution of DTC risk both in the test and validation sets (Supplementary text; Supplementary figure S5). The

 Accepted Manuscript published as ETJ-22-0058.R2. Accepted for publication: 17-Aug-2022

performance of other classifiers was weaker than that of the AdaBoost algorithm (Supplementary figure S3, S6 and S7).

[FIGURE 4]

Finally, the importance of each individual SNP allele in the classification by the trained AdaBoost model was evaluated with the aim of exploring the weight of each individual SNP in the identification of the DTC state (Figure 4B). The most important SNPs, with a relative feature importance greater than 0.6, were rs966423, rs6759952, rs966513, rs7617304, rs3758249, rs10238549, rs4808708, and rs1799814. Interestingly, the top two SNPs, namely rs966423 and rs6759952 were both considered as highly likely to be predictive in the SNP selection phase (see paragraph "SNPs associated with DTC").

It is worth mentioning how some of the SNPs which were deemed "robustly associated with the risk of DTC" or "highly likely DTC risk markers" in the association analysis, such as rs7048394, rs944289 and rs1203952, respectively, are not among the top-ranking in terms of importance in the ML model. This is due to the fact that the prediction of the AdaBoost model is based on the given constellation of all the selected SNPs rather than on the SNPs taken individually. Thus, in this specific classification task the combination of the top-ranking SNPs in Figure 4 might contain enough information, so that some of the SNPs which were strongly associated to DTC risk in the initial association analysis become progressively less important, or even redundant, and do not improve the overall predictive performance further.

DISCUSSION

The present study outlined the potential of a minimal selection of 15 DTC-associated SNPs to confirm the disease predisposition. We re-evaluated the candidate risk *loci* described in the literature as individually associated with DTC. Candidates were verified for their association by consulting the results obtained in a previous GWAS [12]. The most strongly associated SNPs were *ad hoc* genotyped in two independent datasets, accounting for a total of 1131 individuals. The 15 best-performing SNPs were used for the calculation of PRS and wPRS and employed for reconstructing the genetic structure of individuals. Most importantly, we used these SNPs to describe the DTC and healthy control state of individuals using ML. Interestingly, the classification using the AdaBoost algorithm showed fair performance in the test set, with accuracies as high as 64%, as well as in the external validation set, settling at 67% (Table 5). Considering that the ROC AUC is a performance measurement for the classification [39], we may assume to have detected a minimal pool of SNPs consistently contributing to the risk of developing DTC in our Italian dataset, as an example of polygenic disease. This has been further confirmed by the fair confidence of the AdaBoost in classifying the disease state, as highlighted by the PPV value of up to 0.7 on the external validation set. Our results provide a substantial improvement in understanding the impact of genetics on DTC, which until now could be estimated by PRS and could explain only the 11% of the total genetic variability linked to the disease [24,23]. Moreover, the 15 selected SNPs describe the inner genetic structure of sampled individuals when assessed using Bayesian inference from population genetics. We evaluated whether this method would decipher differences between cases and controls, assigning a phenotype-specific genetic footprint to individuals, with a quantitative approach. Individuals were distributed among two subpopulations with different genetic patterns, following the DTC or healthy control state, although a certain degree of admixture was found. This analysis confirmed that the selected SNPs are representative of the genetic signature linked to the disease. When using these SNPs for a ML-based analysis of DTC, we obtained a fair classification power.

 Accepted Manuscript published as ETJ-22-0058.R2. Accepted for publication: 17-Aug-2022

Overall, we found six SNPs within the *FOXE1, PTCSC3-LINC00609, FOXA2* and *DIRC3* genes robustly associated with the risk of DTC or categorized as highly likely risk factors, confirming they are well-established predisposing factors for DTC [8,11,12,14,40,41,13]. SNPs in the *DIRC3* (rs966423, rs6759952) and *FOXE1* (rs3758249) genes were also highly relevant features for DTC risk. Other eight SNPs, such as those falling within the *CYP1A1, NIS-SLC5A5, IL11RA* and *let-7i/LINC01465* genes [42-60], did not replicate formally in the second stage of the study, although they maintained or reinforced the statistical significance of the GWAS in the combined analysis. One SNP falling within the *CYP1A1* gene (rs1799814) was also highly relevant for DTC predisposition. Interestingly, there is relative lack of knowledge on the role of the remaining three genes, i.e. *IMMP2L*, *RARRES1* and *CARD9-SNAPC4* in thyroid cancer. However, since the selected SNPs were associated with DTC in the combined analysis, they could be reasonably involved in the aetiology of the disease. They have been previously involved in the regulation of cell metabolism, estrogen physiology, tumour suppression or progression and autoimmune diseases [61-66]. These three genes are certainly interesting for future studies in connection with DTC in the future. The remaining 137 SNPs were not confirmed in dataset 1, while other 19 SNPs positive in the GWAS were not confirmed in dataset 2. They could have been detected as the consequence of chance findings in underpowered studies published in literature, resulting as false or weakly positive signals. An extensive discussion of these gene SNPs is provided as a supplementary material (Supplementary text).

Considering the 15 most associated SNPs, we also calculated PRS and wPRS, confirming that the disease risk increases together with the number of risk alleles. An OR of 6.9 (95% CI = 5.4-8.8) for the top 10th decile was found based on a 10-SNPs model, including SNPs falling within *PCNX2*, *DIRC3*, *LRRC34*, *EPB41L4A*, *NRG1*, *PTCSC2*, *STN1-SLK*, *PTCSC3 LINC00609*, *MBIP*, and *SMAD3* genes. Our result are in agreement with previous studies concluding that the genetic predisposition to PTC may be resumed by only 10 SNPs, found by wPRS analysis and accounting for between 8 and 11% of the total variability [24,23]. In particular, that study had only three markers in *linkage disequilibrium* with SNPs found herein and lacking the requisites for being selected and run on ML

analysis (Supplementary text). It is correct to specify that previous studies only studied PTC and not DTC. However, we believe the comparison is feasible as PTC accounts for at least 85% of all thyroid cancers. Similar OR values were found in a previous GWAS performed using a 11-SNPs signature [13], that shared only the *DIRC3* gene region with our proposed signature. Taken together, these data indicate that unknown, low-penetrance SNPs contributing to genetic predisposition to DTC may be discovered using different approaches. A recent study on Korean population, found lower ORs (1.46 and 1.56 for unweighted and weighted PRS), but considering only six SNP associated to thyroid cancer [67].

Obviously, our study is limited by the fact of not having considered all the possible SNPs associated with DTC in the literature but only those associated with the DTC risk in the GWAS enrolling subjects of dataset 1. Therefore, classification algorithms relied on the genetic information alone. Data about the exposure to important risk factors, such as ionizing radiation and family history of DTC, were only available for a subset of the study population (not shown). Therefore, they could not be considered in the statistical analyses and for the construction of ML models. Another issue may consist in the ethnicity of datasets used herein, which consists in Italian individuals. The association between these SNPs and the disease would be explored in individuals of different ethnicity. Finally, in case-control studies, proper sample selection is crucial to attain robust disease prediction: individuals recorded as healthy controls might develop DTC even in older age, even if subjects had no thyroid abnormalities at the time of ultrasound analysis. Such individuals should be considered as *spurious negatives*. Similarly, young DTC individuals might have seen the development of the disease following causes beyond the genetic predisposition, such as exposure to ionizing radiation, and might thus represent *spurious positives*. These aspects represent confounding factors when attempting to extrapolate the genetic footprint of the disease used to build ML models. For the reasons listed above, the direct clinical impact of our result is limited. It has yet to be clarified which other genetic markers cover the remaining slice of heritability or predisposition. Then it is necessary to analyze genetics

 Accepted Manuscript published as ETJ-22-0058.R2. Accepted for publication: 17-Aug-2022

together with other environmental risk factors, some of which are difficult to measure, such as exposure to radiation or pollutants.

In conclusion, we described a procedure based on a combined PRS and ML approach that allows a fair description of the case or control state based solely on the individual genetic background. This analysis provided evidence for a new, restricted selection of 15 SNPs associated with the risk of DTC, extending the series previously found using different approaches [24] and further delineating the genetic signature of the disease in our Italian dataset. Further developments might aim to implement and refine the reported methodology with more covariates and might improve the overall accuracy.

Declaration of interest

Authors have no potential conflict of interests.

Funding

KH was supported by the Horizon 2020 Program of the European Union (grant 856620).

Contributions

GB participated to study design, data collection, analysis and interpretation, manuscript writing and revision. CL, EP, FN and MSit were involved in data collection. FT revised te manuscript and was involved in data interpretation and intellectual content. GM, RS and FG was involved in data analysis. AF, KH and RE participated to data collection, interpretation of results and revised the intellectual content of the manuscript. Cristina Romei was involved in data collection and analysis. EAZ and MAD participated to study design, data analysis and interpretation, manuscript writing and revision. MS was involved in interpretation of results and revised the intellectual content of the manuscript. SL and LC conceived the study, participated to study design, data collection, analysis, interpretation and intellectual content, manuscript writing and revision.

Acknowledgments: The authors are grateful to the Italian Ministry of University and Research for supporting the Department of Biomedical, Metabolic, and Neural Sciences (University of Modena and Reggio Emilia, Italy) in the context of the Departments of Excellence Programme. The study was supported by IBSA Institut Biochimique SA, without involvement in study design, collection, analysis, and interpretation of data, writing of the report, nor any restrictions regarding the submission of the report for publication.

 Accepted Manuscript published as ETJ-22-0058.R2. Accepted for publication: 17-Aug-2022

REFERENCES

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A (2018) Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians 68 (6):394-424. doi:10.3322/caac.21492

2. Kitahara CM, Sosa JA (2016) The changing incidence of thyroid cancer. Nature reviews Endocrinology 12 (11):646-653. doi:10.1038/nrendo.2016.110

3. Haugen BR, Alexander EK, Bible KC, Doherty GM, Mandel SJ, Nikiforov YE, Pacini F, Randolph GW, Sawka AM, Schlumberger M, Schuff KG, Sherman SI, Sosa JA, Steward DL, Tuttle RM, Wartofsky L (2016) 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 : official journal of the American Thyroid Association 26 (1):1-133. doi:10.1089/thy.2015.0020

4. Cabanillas ME, McFadden DG, Durante C (2016) Thyroid cancer. Lancet (London, England) 388 (10061):2783-2795. doi:10.1016/s0140-6736(16)30172-6

5. Fallah M, Pukkala E, Tryggvadottir L, Olsen JH, Tretli S, Sundquist K, Hemminki K (2013) Risk of thyroid cancer in first-degree relatives of patients with non-medullary thyroid cancer by histology type and age at diagnosis: a joint study from five Nordic countries. Journal of medical genetics 50 (6):373-382. doi:10.1136/jmedgenet-2012-101412

6. Hemminki K, Sundquist J, Lorenzo Bermejo J (2008) Familial risks for cancer as the basis for evidence-based clinical referral and counseling. The oncologist 13 (3):239-247. doi:10.1634/theoncologist.2007-0242

7. Saenko VA, Rogounovitch TI (2018) Genetic Polymorphism Predisposing to Differentiated Thyroid Cancer: A Review of Major Findings of the Genome-Wide Association Studies. Endocrinology and metabolism (Seoul, Korea) 33 (2):164-174. doi:10.3803/EnM.2018.33.2.164

8. Gudmundsson J, Sulem P, Gudbjartsson DF, Jonasson JG, Sigurdsson A, Bergthorsson JT, He H, Blondal T, Geller F, Jakobsdottir M, Magnusdottir DN, Matthiasdottir S, Stacey SN, Skarphedinsson

OB, Helgadottir H, Li W, Nagy R, Aguillo E, Faure E, Prats E, Saez B, Martinez M, Eyjolfsson GI, Bjornsdottir US, Holm H, Kristjansson K, Frigge ML, Kristvinsson H, Gulcher JR, Jonsson T, Rafnar T, Hjartarsson H, Mayordomo JI, de la Chapelle A, Hrafnkelsson J, Thorsteinsdottir U, Kong A, Stefansson K (2009) Common variants on 9q22.33 and 14q13.3 predispose to thyroid cancer in European populations. Nature genetics 41 (4):460-464. doi:10.1038/ng.339

9. Takahashi M, Saenko VA, Rogounovitch TI, Kawaguchi T, Drozd VM, Takigawa-Imamura H, Akulevich NM, Ratanajaraya C, Mitsutake N, Takamura N, Danilova LI, Lushchik ML, Demidchik YE, Heath S, Yamada R, Lathrop M, Matsuda F, Yamashita S (2010) The FOXE1 locus is a major genetic determinant for radiation-related thyroid carcinoma in Chernobyl. Human molecular genetics 19 (12):2516-2523. doi:10.1093/hmg/ddq123

10. Jendrzejewski J, He H, Radomska HS, Li W, Tomsic J, Liyanarachchi S, Davuluri RV, Nagy R, de la Chapelle A (2012) The polymorphism rs944289 predisposes to papillary thyroid carcinoma through a large intergenic noncoding RNA gene of tumor suppressor type. Proceedings of the National Academy of Sciences of the United States of America 109 (22):8646-8651. doi:10.1073/pnas.1205654109

11. Gudmundsson J, Sulem P, Gudbjartsson DF, Jonasson JG, Masson G, He H, Jonasdottir A, Sigurdsson A, Stacey SN, Johannsdottir H, Helgadottir HT, Li W, Nagy R, Ringel MD, Kloos RT, de Visser MC, Plantinga TS, den Heijer M, Aguillo E, Panadero A, Prats E, Garcia-Castaño A, De Juan A, Rivera F, Walters GB, Bjarnason H, Tryggvadottir L, Eyjolfsson GI, Bjornsdottir US, Holm H, Olafsson I, Kristjansson K, Kristvinsson H, Magnusson OT, Thorleifsson G, Gulcher JR, Kong A, Kiemeney LA, Jonsson T, Hjartarson H, Mayordomo JI, Netea-Maier RT, de la Chapelle A, Hrafnkelsson J, Thorsteinsdottir U, Rafnar T, Stefansson K (2012) Discovery of common variants associated with low TSH levels and thyroid cancer risk. Nature genetics 44 (3):319-322. doi:10.1038/ng.1046

12. Köhler A, Chen B, Gemignani F, Elisei R, Romei C, Figlioli G, Cipollini M, Cristaudo A, Bambi F, Hoffmann P, Herms S, Kalemba M, Kula D, Harris S, Broderick P, Houlston R, Pastor S, Marcos

 Accepted Manuscript published as ETJ-22-0058.R2. Accepted for publication: 17-Aug-2022

R, Velázquez A, Jarzab B, Hemminki K, Landi S, Försti A (2013) Genome-wide association study on differentiated thyroid cancer. The Journal of clinical endocrinology and metabolism 98 (10):E1674-1681. doi:10.1210/jc.2013-1941

13. Figlioli G, Köhler A, Chen B, Elisei R, Romei C, Cipollini M, Cristaudo A, Bambi F, Paolicchi E, Hoffmann P, Herms S, Kalemba M, Kula D, Pastor S, Marcos R, Velázquez A, Jarząb B, Landi S, Hemminki K, Försti A, Gemignani F (2014) Novel genome-wide association study-based candidate loci for differentiated thyroid cancer risk. The Journal of clinical endocrinology and metabolism 99 (10):E2084-2092. doi:10.1210/jc.2014-1734

14. Son HY, Hwangbo Y, Yoo SK, Im SW, Yang SD, Kwak SJ, Park MS, Kwak SH, Cho SW, Ryu JS, Kim J, Jung YS, Kim TH, Kim SJ, Lee KE, Park DJ, Cho NH, Sung J, Seo JS, Lee EK, Park YJ, Kim JI (2017) Genome-wide association and expression quantitative trait loci studies identify multiple susceptibility loci for thyroid cancer. Nature communications 8:15966. doi:10.1038/ncomms15966

15. Gudmundsson J, Thorleifsson G, Sigurdsson JK, Stefansdottir L, Jonasson JG, Gudjonsson SA, Gudbjartsson DF, Masson G, Johannsdottir H, Halldorsson GH, Stacey SN, Helgason H, Sulem P, Senter L, He H, Liyanarachchi S, Ringel MD, Aguillo E, Panadero A, Prats E, Garcia-Castano A, De Juan A, Rivera F, Xu L, Kiemeney LA, Eyjolfsson GI, Sigurdardottir O, Olafsson I, Kristvinsson H, Netea-Maier RT, Jonsson T, Mayordomo JI, Plantinga TS, Hjartarson H, Hrafnkelsson J, Sturgis EM, Thorsteinsdottir U, Rafnar T, de la Chapelle A, Stefansson K (2017) A genome-wide association study yields five novel thyroid cancer risk loci. Nature communications 8:14517. doi:10.1038/ncomms14517

16. Yanes T, Young MA, Meiser B, James PA (2020) Clinical applications of polygenic breast cancer risk: a critical review and perspectives of an emerging field. Breast cancer research : BCR 22 (1):21. doi:10.1186/s13058-020-01260-3

17. Lambert SA, Abraham G, Inouye M (2019) Towards clinical utility of polygenic risk scores. Human molecular genetics 28 (R2):R133-r142. doi:10.1093/hmg/ddz187

18. Galeotti AA, Gentiluomo M, Rizzato C, Obazee O, Neoptolemos JP, Pasquali C, Nentwich M, Cavestro GM, Pezzilli R, Greenhalf W, Holleczek B, Schroeder C, Schöttker B, Ivanauskas A, Ginocchi L, Key TJ, Hegyi P, Archibugi L, Darvasi E, Basso D, Sperti C, Bijlsma MF, Palmieri O, Hlavac V, Talar-Wojnarowska R, Mohelnikova-Duchonova B, Hackert T, Vashist Y, Strouhal O, van Laarhoven H, Tavano F, Lovecek M, Dervenis C, Izbéki F, Padoan A, Małecka-Panas E, Maiello E, Vanella G, Capurso G, Izbicki JR, Theodoropoulos GE, Jamroziak K, Katzke V, Kaaks R, Mambrini A, Papanikolaou IS, Szmola R, Szentesi A, Kupcinskas J, Bursi S, Costello E, Boggi U, Milanetto AC, Landi S, Gazouli M, Vodickova L, Soucek P, Gioffreda D, Gemignani F, Brenner H, Strobel O, Büchler M, Vodicka P, Paiella S, Canzian F, Campa D (2020) Polygenic and multifactorial scores for pancreatic ductal adenocarcinoma risk prediction. Journal of medical genetics. doi:10.1136/jmedgenet-2020-106961

19. Lindström S, Schumacher FR, Cox D, Travis RC, Albanes D, Allen NE, Andriole G, Berndt SI, Boeing H, Bueno-de-Mesquita HB, Crawford ED, Diver WR, Gaziano JM, Giles GG, Giovannucci E, Gonzalez CA, Henderson B, Hunter DJ, Johansson M, Kolonel LN, Ma J, Le Marchand L, Pala V, Stampfer M, Stram DO, Thun MJ, Tjonneland A, Trichopoulos D, Virtamo J, Weinstein SJ, Willett WC, Yeager M, Hayes RB, Severi G, Haiman CA, Chanock SJ, Kraft P (2012) Common genetic variants in prostate cancer risk prediction--results from the NCI Breast and Prostate Cancer Cohort Consortium (BPC3). Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology 21 (3):437-444. doi:10.1158/1055-9965.Epi-11-1038

20. Hüsing A, Canzian F, Beckmann L, Garcia-Closas M, Diver WR, Thun MJ, Berg CD, Hoover RN, Ziegler RG, Figueroa JD, Isaacs C, Olsen A, Viallon V, Boeing H, Masala G, Trichopoulos D, Peeters PH, Lund E, Ardanaz E, Khaw KT, Lenner P, Kolonel LN, Stram DO, Le Marchand L, McCarty CA, Buring JE, Lee IM, Zhang S, Lindström S, Hankinson SE, Riboli E, Hunter DJ, Henderson BE, Chanock SJ, Haiman CA, Kraft P, Kaaks R (2012) Prediction of breast cancer risk

 Accepted Manuscript published as ETJ-22-0058.R2. Accepted for publication: 17-Aug-2022

by genetic risk factors, overall and by hormone receptor status. Journal of medical genetics 49 (9):601-608. doi:10.1136/jmedgenet-2011-100716

21. Mavaddat N, Pharoah PD, Michailidou K, Tyrer J, Brook MN, Bolla MK, Wang Q, Dennis J, Dunning AM, Shah M, Luben R, Brown J, Bojesen SE, Nordestgaard BG, Nielsen SF, Flyger H, Czene K, Darabi H, Eriksson M, Peto J, Dos-Santos-Silva I, Dudbridge F, Johnson N, Schmidt MK, Broeks A, Verhoef S, Rutgers EJ, Swerdlow A, Ashworth A, Orr N, Schoemaker MJ, Figueroa J, Chanock SJ, Brinton L, Lissowska J, Couch FJ, Olson JE, Vachon C, Pankratz VS, Lambrechts D, Wildiers H, Van Ongeval C, van Limbergen E, Kristensen V, Grenaker Alnæs G, Nord S, Borresen-Dale AL, Nevanlinna H, Muranen TA, Aittomäki K, Blomqvist C, Chang-Claude J, Rudolph A, Seibold P, Flesch-Janys D, Fasching PA, Haeberle L, Ekici AB, Beckmann MW, Burwinkel B, Marme F, Schneeweiss A, Sohn C, Trentham-Dietz A, Newcomb P, Titus L, Egan KM, Hunter DJ, Lindstrom S, Tamimi RM, Kraft P, Rahman N, Turnbull C, Renwick A, Seal S, Li J, Liu J, Humphreys K, Benitez J, Pilar Zamora M, Arias Perez JI, Menéndez P, Jakubowska A, Lubinski J, Jaworska-Bieniek K, Durda K, Bogdanova NV, Antonenkova NN, Dörk T, Anton-Culver H, Neuhausen SL, Ziogas A, Bernstein L, Devilee P, Tollenaar RA, Seynaeve C, van Asperen CJ, Cox A, Cross SS, Reed MW, Khusnutdinova E, Bermisheva M, Prokofyeva D, Takhirova Z, Meindl A, Schmutzler RK, Sutter C, Yang R, Schürmann P, Bremer M, Christiansen H, Park-Simon TW, Hillemanns P, Guénel P, Truong T, Menegaux F, Sanchez M, Radice P, Peterlongo P, Manoukian S, Pensotti V, Hopper JL, Tsimiklis H, Apicella C, Southey MC, Brauch H, Brüning T, Ko YD, Sigurdson AJ, Doody MM, Hamann U, Torres D, Ulmer HU, Försti A, Sawyer EJ, Tomlinson I, Kerin MJ, Miller N, Andrulis IL, Knight JA, Glendon G, Marie Mulligan A, Chenevix-Trench G, Balleine R, Giles GG, Milne RL, McLean C, Lindblom A, Margolin S, Haiman CA, Henderson BE, Schumacher F, Le Marchand L, Eilber U, Wang-Gohrke S, Hooning MJ, Hollestelle A, van den Ouweland AM, Koppert LB, Carpenter J, Clarke C, Scott R, Mannermaa A, Kataja V, Kosma VM, Hartikainen JM, Brenner H, Arndt V, Stegmaier C, Karina Dieffenbach A, Winqvist R, Pylkäs K, Jukkola-Vuorinen A, Grip M, Offit K, Vijai J, Robson M, Rau-Murthy R, Dwek M, Swann R, Annie

Perkins K, Goldberg MS, Labrèche F, Dumont M, Eccles DM, Tapper WJ, Rafiq S, John EM, Whittemore AS, Slager S, Yannoukakos D, Toland AE, Yao S, Zheng W, Halverson SL, González-Neira A, Pita G, Rosario Alonso M, Álvarez N, Herrero D, Tessier DC, Vincent D, Bacot F, Luccarini C, Baynes C, Ahmed S, Maranian M, Healey CS, Simard J, Hall P, Easton DF, Garcia-Closas M (2015) Prediction of breast cancer risk based on profiling with common genetic variants. Journal of the National Cancer Institute 107 (5). doi:10.1093/jnci/djv036

22. Hüsing A, Dossus L, Ferrari P, Tjønneland A, Hansen L, Fagherazzi G, Baglietto L, Schock H, Chang-Claude J, Boeing H, Steffen A, Trichopoulou A, Bamia C, Katsoulis M, Krogh V, Palli D, Panico S, Onland-Moret NC, Peeters PH, Bueno-de-Mesquita HB, Weiderpass E, Gram IT, Ardanaz E, Obón-Santacana M, Navarro C, Sánchez-Cantalejo E, Etxezarreta N, Allen NE, Khaw KT, Wareham N, Rinaldi S, Romieu I, Merritt MA, Gunter M, Riboli E, Kaaks R (2016) An epidemiological model for prediction of endometrial cancer risk in Europe. European journal of epidemiology 31 (1):51-60. doi:10.1007/s10654-015-0030-9

23. Liyanarachchi S, Wojcicka A, Li W, Czetwertynska M, Stachlewska E, Nagy R, Hoag K, Wen B, Ploski R, Ringel MD, Kozłowicz-Gudzinska I, Gierlikowski W, Jazdzewski K, He H, de la Chapelle A (2013) Cumulative risk impact of five genetic variants associated with papillary thyroid carcinoma. Thyroid : official journal of the American Thyroid Association 23 (12):1532-1540. doi:10.1089/thy.2013.0102

24. Liyanarachchi S, Gudmundsson J, Ferkingstad E, He H, Jonasson JG, Tragante V, Asselbergs FW, Xu L, Kiemeney LA, Netea-Maier RT, Mayordomo JI, Plantinga TS, Hjartarson H, Hrafnkelsson J, Sturgis EM, Brock P, Nabhan F, Thorleifsson G, Ringel MD, Stefansson K, de la Chapelle A (2020) Assessing thyroid cancer risk using polygenic risk scores. Proceedings of the National Academy of Sciences of the United States of America 117 (11):5997-6002. doi:10.1073/pnas.1919976117

 Accepted Manuscript published as ETJ-22-0058.R2. Accepted for publication: 17-Aug-2022

25. Kim BJ, Kim SH (2018) Prediction of inherited genomic susceptibility to 20 common cancer types by a supervised machine-learning method. Proceedings of the National Academy of Sciences of the United States of America 115 (6):1322-1327. doi:10.1073/pnas.1717960115

26. Ge T, Chen CY, Ni Y, Feng YA, Smoller JW (2019) Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nature communications 10 (1):1776. doi:10.1038/s41467-019-09718-5

27. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW (2002) Genetic structure of human populations. Science (New York, NY) 298 (5602):2381-2385. doi:10.1126/science.1078311

28. Luyapan J, Ji X, Li S, Xiao X, Zhu D, Duell EJ, Christiani DC, Schabath MB, Arnold SM, Zienolddiny S, Brunnström H, Melander O, Thornquist MD, MacKenzie TA, Amos CI, Gui J (2020) A new efficient method to detect genetic interactions for lung cancer GWAS. BMC medical genomics 13 (1):162. doi:10.1186/s12920-020-00807-9

29. Hu YJ, Lin DY (2010) Analysis of untyped SNPs: maximum likelihood and imputation methods. Genetic epidemiology 34 (8):803-815. doi:10.1002/gepi.20527

30. Hothorn LA, Hothorn T (2009) Order-restricted scores test for the evaluation of population-based case-control studies when the genetic model is unknown. Biometrical journal Biometrische Zeitschrift 51 (4):659-669. doi:10.1002/bimj.200800203

31. Moldovan A, Waldman YY, Brandes N, Linial M (2021) Body Mass Index and Birth Weight Improve Polygenic Risk Score for Type 2 Diabetes. Journal of personalized medicine 11 (6). doi:10.3390/jpm11060582

32. Casarini L, Brigante G (2014) The polycystic ovary syndrome evolutionary paradox: a genome-wide association studies-based, in silico, evolutionary explanation. The Journal of clinical endocrinology and metabolism 99 (11):E2412-2420. doi:10.1210/jc.2014-2703

33. Ng AY, Jordan MI (2001) On discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes. Paper presented at the Proceedings of the 14th International Conference

on Neural Information Processing Systems: Natural and Synthetic, Vancouver, British Columbia, Canada,

34. Friedman JH (2001) Greedy function approximation: A gradient boosting machine. The Annals of Statistics 29 (5):1189-1232, 1144

35. Freund Y, Schapire RE A desicion-theoretic generalization of on-line learning and an application to boosting. In, Berlin, Heidelberg, 1995. Computational Learning Theory. Springer Berlin Heidelberg, pp 23-37

36. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É (2011) Scikit-learn: Machine Learning in Python. J Mach Learn Res 12 (null):2825–2830

37. Freund Y, Schapire RE (1996) Experiments with a new boosting algorithm. Paper presented at the Proceedings of the Thirteenth International Conference on International Conference on Machine Learning, Bari, Italy,

38. Platt J (1999) Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers 10 (3):61-74

39. Lasko TA, Bhagwat JG, Zou KH, Ohno-Machado L (2005) The use of receiver operating characteristic curves in biomedical informatics. Journal of biomedical informatics 38 (5):404-415. doi:10.1016/j.jbi.2005.02.008

40. Wang YL, Feng SH, Guo SC, Wei WJ, Li DS, Wang Y, Wang X, Wang ZY, Ma YY, Jin L, Ji QH, Wang JC (2013) Confirmation of papillary thyroid cancer susceptibility loci identified by genome-wide association studies of chromosomes 14q13, 9q22, 2q35 and 8p12 in a Chinese population. Journal of medical genetics 50 (10):689-695. doi:10.1136/jmedgenet-2013-101687

41. Kim HS, Kim DH, Kim JY, Jeoung NH, Lee IK, Bong JG, Jung ED (2010) Microarray analysis of papillary thyroid cancers in Korean. The Korean journal of internal medicine 25 (4):399-407. doi:10.3904/kjim.2010.25.4.399

 Accepted Manuscript published as ETJ-22-0058.R2. Accepted for publication: 17-Aug-2022

42. Figlioli G, Elisei R, Romei C, Melaiu O, Cipollini M, Bambi F, Chen B, Köhler A, Cristaudo A, Hemminki K, Gemignani F, Försti A, Landi S (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 : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology 25 (4):700-713. doi:10.1158/1055-9965.Epi-15-0652

43. Siraj AK, Ibrahim M, Al-Rasheed M, Abubaker J, Bu R, Siddiqui SU, Al-Dayel F, Al-Sanea O, Al-Nuaim A, Uddin S, Al-Kuraya K (2008) Polymorphisms of selected xenobiotic genes contribute to the development of papillary thyroid cancer susceptibility in Middle Eastern population. BMC medical genetics 9:61. doi:10.1186/1471-2350-9-61

44. Irmiakova AR, Kochetova OV, Gaĭnullina MK, Sivochalova OV, Viktorova TV (2012) [Association of polymorph variants of CYP1A2 and CYP1A1 genes with reproductive and thyroid diseases in female workers of petrochemical industry]. Meditsina truda i promyshlennaia ekologiia (5):41-48

45. Bufalo NE, Leite JL, Guilhen AC, Morari EC, Granja F, Assumpcao LV, Ward LS (2006) Smoking and susceptibility to thyroid cancer: an inverse association with CYP1A1 allelic variants. Endocrine-related cancer 13 (4):1185-1193. doi:10.1677/erc-06-0002

46. GallegosVargas J, SanchezRoldan J, RonquilloSanchez M, Carmona Aparicio L, FlorianoSanchez E, CardenasRodriguez N (2016) Gene Expression of CYP1A1 and its Possible Clinical Application in Thyroid Cancer Cases. Asian Pacific journal of cancer prevention : APJCP 17 (7):3477-3482

47. de Morais RM, Sobrinho AB, de Souza Silva CM, de Oliveira JR, da Silva ICR, de Toledo Nóbrega O (2018) The Role of the NIS (SLC5A5) Gene in Papillary Thyroid Cancer: A Systematic Review. International journal of endocrinology 2018:9128754. doi:10.1155/2018/9128754

48. Heinrich PC, Behrmann I, Müller-Newen G, Schaper F, Graeve L (1998) Interleukin-6-type cytokine signalling through the gp130/Jak/STAT pathway. The Biochemical journal 334 (Pt 2) (Pt 2):297-314. doi:10.1042/bj3340297

49. Katoh M, Katoh M (2007) STAT3-induced WNT5A signaling loop in embryonic stem cells, adult normal tissues, chronic persistent inflammation, rheumatoid arthritis and cancer (Review). International journal of molecular medicine 19 (2):273-278

50. Hanavadi S, Martin TA, Watkins G, Mansel RE, Jiang WG (2006) Expression of interleukin 11 and its receptor and their prognostic value in human breast cancer. Annals of surgical oncology 13 (6):802-808. doi:10.1245/aso.2006.05.028

51. Goseki N, Koike M, Yoshida M (1992) Histopathologic characteristics of early stage esophageal carcinoma. A comparative study with gastric carcinoma. Cancer 69 (5):1088-1093. doi:10.1002/cncr.2820690503

52. Yamazumi K, Nakayama T, Kusaba T, Wen CY, Yoshizaki A, Yakata Y, Nagayasu T, Sekine I (2006) Expression of interleukin-11 and interleukin-11 receptor alpha in human colorectal adenocarcinoma; immunohistochemical analyses and correlation with clinicopathological factors. World journal of gastroenterology 12 (2):317-321. doi:10.3748/wjg.v12.i2.317

53. Vogelstein B, Fearon ER, Hamilton SR, Kern SE, Preisinger AC, Leppert M, Nakamura Y, White R, Smits AM, Bos JL (1988) Genetic alterations during colorectal-tumor development. The New England journal of medicine 319 (9):525-532. doi:10.1056/nejm198809013190901

54. Eun YG, Shin IH, Kim MJ, Chung JH, Song JY, Kwon KH (2012) Associations between promoter polymorphism -106A/G of interleukin-11 receptor alpha and papillary thyroid cancer in Korean population. Surgery 151 (2):323-329. doi:10.1016/j.surg.2011.07.014

55. Lin P, Guo YN, Shi L, Li XJ, Yang H, He Y, Li Q, Dang YW, Wei KL, Chen G (2019) Development of a prognostic index based on an immunogenomic landscape analysis of papillary thyroid cancer. Aging 11 (2):480-500. doi:10.18632/aging.101754

 Accepted Manuscript published as ETJ-22-0058.R2. Accepted for publication: 17-Aug-2022

56. Zhong Z, Hu Z, Jiang Y, Sun R, Chen X, Chu H, Zeng M, Sun C (2016) Interleukin-11 promotes epithelial-mesenchymal transition in anaplastic thyroid carcinoma cells through PI3K/Akt/GSK3β signaling pathway activation. Oncotarget 7 (37):59652-59663. doi:10.18632/oncotarget.10831

57. Wang Y, Wei T, Xiong J, Chen P, Wang X, Zhang L, Gao L, Zhu J (2015) Association Between Genetic Polymorphisms in the Promoter Regions of Let-7 and Risk of Papillary Thyroid Carcinoma: A Case-Control Study. Medicine 94 (43):e1879. doi:10.1097/md.0000000000001879

58. Perdas E, Stawski R, Kaczka K, Zubrzycka M (2020) Analysis of Let-7 Family miRNA in Plasma as Potential Predictive Biomarkers of Diagnosis for Papillary Thyroid Cancer. Diagnostics (Basel, Switzerland) 10 (3). doi:10.3390/diagnostics10030130

59. Li M, Song Q, Li H, Lou Y, Wang L (2015) Circulating miR-25-3p and miR-451a May Be Potential Biomarkers for the Diagnosis of Papillary Thyroid Carcinoma. PloS one 10 (7):e0132403. doi:10.1371/journal.pone.0132403

60. Perdas E, Stawski R, Nowak D, Zubrzycka M (2016) The Role of miRNA in Papillary Thyroid Cancer in the Context of miRNA Let-7 Family. International journal of molecular sciences 17 (6). doi:10.3390/ijms17060909

61. Yuan L, Zhai L, Qian L, Huang D, Ding Y, Xiang H, Liu X, Thompson JW, Liu J, He YH, Chen XQ, Hu J, Kong QP, Tan M, Wang XF (2018) Switching off IMMP2L signaling drives senescence via simultaneous metabolic alteration and blockage of cell death. Cell research 28 (6):625-643. doi:10.1038/s41422-018-0043-5

62. Kloth M, Goering W, Ribarska T, Arsov C, Sorensen KD, Schulz WA (2012) The SNP rs6441224 influences transcriptional activity and prognostically relevant hypermethylation of RARRES1 in prostate cancer. International journal of cancer 131 (6):E897-904. doi:10.1002/ijc.27628

63. Yanatatsaneejit P, Chalermchai T, Kerekhanjanarong V, Shotelersuk K, Supiyaphun P, Mutirangura A, Sriuranpong V (2008) Promoter hypermethylation of CCNA1, RARRES1, and HRASLS3 in nasopharyngeal carcinoma. Oral oncology 44 (4):400-406. doi:10.1016/j.oraloncology.2007.05.008

64. Wilson CL, Sims AH, Howell A, Miller CJ, Clarke RB (2006) Effects of oestrogen on gene expression in epithelium and stroma of normal human breast tissue. Endocrine-related cancer 13 (2):617-628. doi:10.1677/erc.1.01165

65. Qu J, Liu L, Xu Q, Ren J, Xu Z, Dou H, Shen S, Hou Y, Mou Y, Wang T (2019) CARD9 prevents lung cancer development by suppressing the expansion of myeloid-derived suppressor cells and IDO production. International journal of cancer 145 (8):2225-2237. doi:10.1002/ijc.32355

66. Németh T, Futosi K, Weisinger J, Csorba K, Sitaru C, Ruland J, Mócsai A (2014) A8.25 CARD9 mediates autoantibody-induced autoimmune diseases by linking the SYK tyrosine kinase to CHEMOKINE production. Annals of the Rheumatic Diseases 73 (Suppl 1):A86-A86. doi:10.1136/annrheumdis-2013-205124.199

67. Hoang T, Nguyen Ngoc Q, Lee J, Lee EK, Hwangbo Y, Kim J (2021) Evaluation of modifiable factors and polygenic risk score in thyroid cancer. Endocrine-related cancer 28 (7):481-494. doi:10.1530/erc-21-0078

 Accepted Manuscript published as ETJ-22-0058.R2. Accepted for publication: 17-Aug-2022

FIGURE LEGENDS

Figure 1. Datasets and project's pipeline. A) Summary of dataset composition, highlighting the progressive refinement of the SNP selection process. Dataset 1 SNPs were extracted from a GWAS [12], while datasets 2 and 3 SNPs were genotyped ad hoc for potentially informative SNPs. The 34 SNPs significantly associated with DTC in dataset 1 were genotyped ad hoc and checked for relevance in the independent dataset 2. Then, 15 SNPs highly associated with DTC in both datasets 1 and 2 were further genotyped ad hoc in the independent dataset 3. B) Procedure for statistical SNP discovery and subsequent ML implementation. After SNPs selection, we tested the capability of the 15 selected SNPs to provide a DTC genetic signature in the merged datasets 1, 2 and 3 with Bayesian statistics for population genetics. Then, Machine Learning methods were run to confirm the case/control state of individuals, using the selected 15 SNPs as input variables. An extended dataset was built by merging the two largest datasets (1 and 3), to obtain a pool of randomly chosen "training" (80% of the merged dataset) and "testing data" (20%). After finding the most effective ML algorithms, a validation analysis was set on the dataset 2. Shaded colours highlight involved datasets. Yellow = dataset 1; green = dataset 2; light-blue = dataset 3.

Figure 2. SNP selection. A) Criteria used for SNP selection. B) SNP subsets. Among the 171 SNPs selected from the literature (Supplementary table S1), only 34 were associated with DTC in the dataset 1 ($p<0.05$; Supplementary table S2) and genotyped in the dataset 2. 15 SNPs were finally selected from datasets 2 as variables for ML analysis and genotyped in the dataset 3 (bold). Panels A and B have matched colours and letters.

Figure 3. DTC-associated genetic structure of cases and healthy controls. Bar plot was calculated by the STRUCTURE software in the merged datasets 1, 2 and 3. Each individual is represented by a vertical line, in which colours indicate the contribution of each of the k=5 components to the individual genetic background. Cases and controls were ordered for graphical reasons, showing different genetic profiles at a glance, although indicating a certain degree of admixture.

Figure 4. Results from the ML-based DTC prediction and SNP relative importance. A) ROC curves obtained on all datasets with the AdaBoost model. Dashed line represents random choice. B) Relative feature importance of all variables (SNPs) in the AdaBoost model. Data normalized to most important feature. Suffix "_2" indicates the second allele. Feature importance is calculated as an average over the individual classifiers used for probability calibration.

TABLE LEGENDS

Table 1. Summary of training, testing and validation ML datasets.

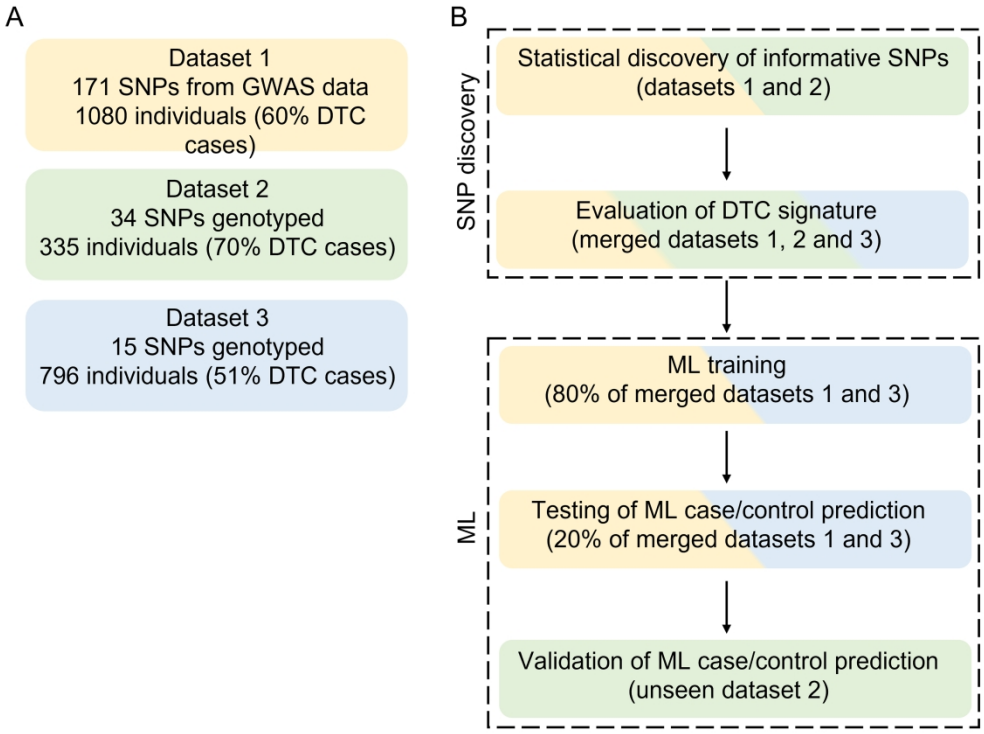
Table 2. Characteristics of study population.

Table 3. List of the 15 SNPs associated with DTC in datasets 1 and 2.

Table 4. Odds ratio estimates for the 15-SNPs PRS quintiles. DTC state obtained in the three merged datasets was considered, using the bottom quintile (0 to 20%) as the reference group. The multivariate logistic regression model included the adjustment of ORs for age, BMI, and gender. wPRS: weighted polygenic risk score; PRS: unweighted polygenic risk score.

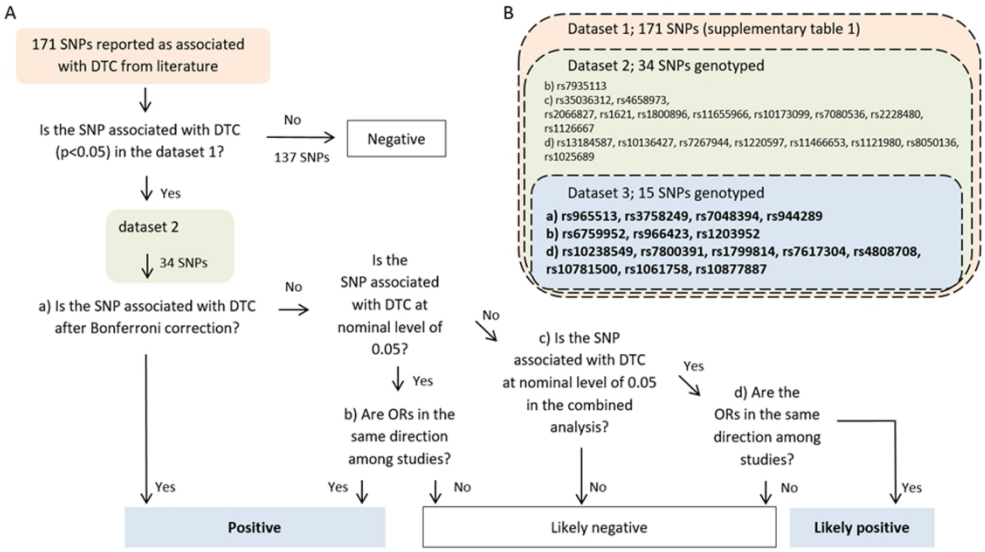
Table 5. Classification metrics of AdaBoost classifier on all datasets.

Accepted Manuscript published as ETJ-22-0058.R2. Accepted for publication: 17-Aug-2022



Datasets and project's pipeline. A) Summary of dataset composition, highlighting the progressive refinement of the SNP selection process. Dataset 1 SNPs were extracted from a GWAS [12], while datasets 2 and 3 SNPs were genotyped ad hoc for potentially informative SNPs. The 34 SNPs significantly associated with DTC in dataset 1 were genotyped ad hoc and checked for relevance in the independent dataset 2. Then, 15 SNPs highly associated with DTC in both datasets 1 and 2 were further genotyped ad hoc in the independent dataset 3. B) Procedure for statistical SNP discovery and subsequent ML implementation. After SNPs selection, we tested the capability of the 15 selected SNPs to provide a DTC genetic signature in the merged datasets 1, 2 and 3 with Bayesian statistics for population genetics. Then, Machine Learning methods were run to confirm the case/control state of individuals, using the selected 15 SNPs as input variables. An extended dataset was built by merging the two largest datasets (1 and 3), to obtain a pool of randomly chosen "training" (80% of the merged dataset) and "testing data" (20%). After finding the most effective ML algorithms, a validation analysis was set on the dataset 2.

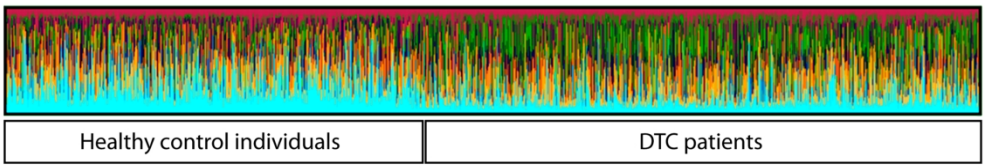
140x103mm (1000 x 1000 DPI)



SNP selection. A) Criteria used for SNP selection. B) SNP subsets. Among the 171 SNPs selected from the literature (Supplementary table S1), only 34 were associated with DTC in the dataset 1 ($p<0.05$; Supplementary table S2) and genotyped in the dataset 2. 15 SNPs were finally selected from datasets 2 as variables for ML analysis and genotyped in the dataset 3 (bold). Panels A and B have matched colours and letters.

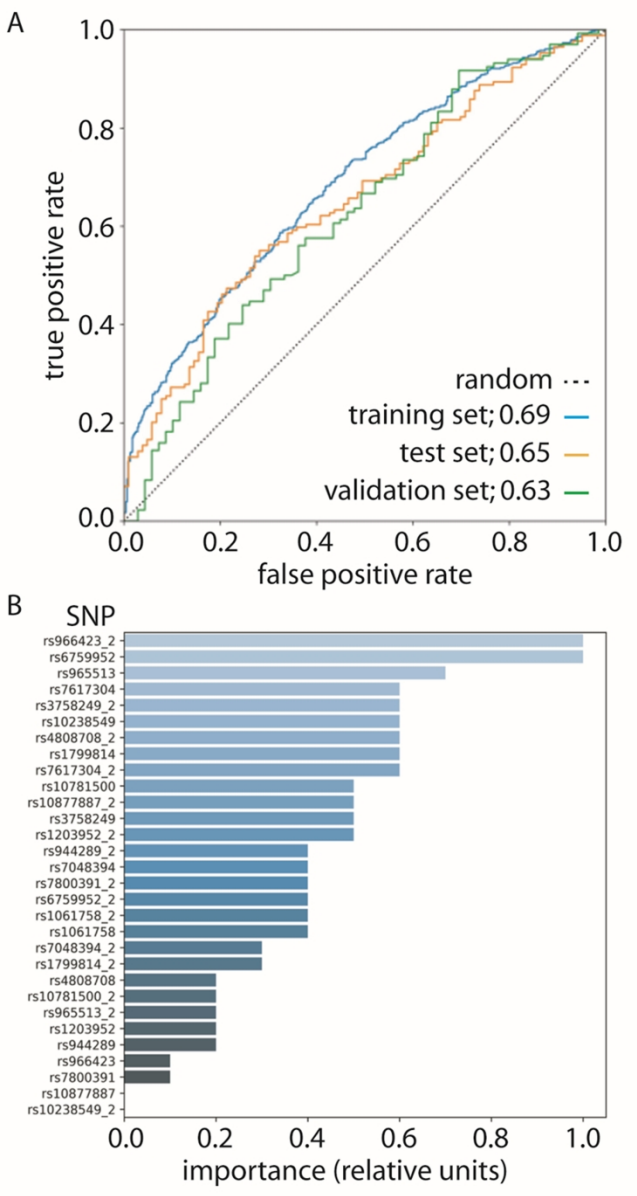
179x102mm (800 x 800 DPI)

Accepted Manuscript published as ETJ-22-0058.R2. Accepted for publication: 17-Aug-2022



DTC-associated genetic structure of cases and healthy controls. Bar plot was calculated by the STRUCTURE software in the merged datasets 1, 2 and 3. Each individual is represented by a vertical line, in which colours indicate the contribution of each of the k=5 components to the individual genetic background. Cases and controls were ordered for graphical reasons, showing different genetic profiles at a glance, although indicating a certain degree of admixture.

189x32mm (1200 x 1200 DPI)



Results from the ML-based DTC prediction and SNP relative importance. A) ROC curves obtained on all datasets with the AdaBoost model. Dashed line represents random choice. B) Relative feature importance of all variables (SNPs) in the AdaBoost model. Data normalized to most important feature. Suffix "_2" indicates the second allele. Feature importance is calculated as an average over the individual classifiers used for probability calibration.

90x169mm (1200 x 1200 DPI)

 Accepted Manuscript published as ETJ-22-0058.R2. Accepted for publication: 17-Aug-2022

ML dataset	Origin of dataset*	# Individuals	Cases (%)	Controls (%)
Training	80% datasets 1 + 3	1086	58.2	41.8
Testing	20% datasets 1 + 3	272	62.1	37.9
Validation	100% dataset 2	201	65.7	34.3

* Summary of datasets after removing individuals with missing data in the genotype (% cases; % controls): dataset 1 = 949 (59.5; 40.5); dataset 2 = 201 (65.7; 34.3); dataset 3 = 409 (57.7; 42.3)

Table 1. Summary of training, testing and validation ML datasets.

	Dataset 1		Dataset 2		Dataset 3	
	Cases (n=649)	Controls (n=431)	Cases (n=234)	Controls (n=101)	Cases (n=404)	Controls (n=392)
Females (%)	507 (78%)	320 (74%)	167 (71%)	61 (60%)	287 (71%)	243 (62%)
Age (years)	37.8±0.85	46.8±0.97	49.7±14.0	43.7±11.4	44.8±12.7	43.8±9.6
Weight (Kg)	71.0±1.31	70.4±1.37	74.5±15.0	71.3±16.1	79.3±17.9	69.2±14.5
BMI (Kg/m^2)	25.3±0.39	25.2±0.38	26.9±4.9	26.1±5.0	27.5±4.7	23.9±3.7

Values are expressed as number and percentages (%) or average and standard error. BMI = body mass index.

Table 2. Characteristics of study population.

Accepted Manuscript published as ETJ-22-0058.R2. Accepted for publication: 17-Aug-2022

SNP ID	Genomic location	Gene	Description
rs965513	chr9:97793827	*PTCSC2/FOXE1*	Papillary thyroid carcinoma susceptibility candidate 2/Forkhead box E1
rs3758249	chr9:97851858	*PTCSC2/FOXE1*	Papillary thyroid carcinoma susceptibility candidate 2/Forkhead box E1
rs7048394	chr9:97843151	*PTCSC2/FOXE1*	Papillary thyroid carcinoma susceptibility candidate 2/Forkhead box E1
rs944289	chr14:36180040	*PTCSC3/LINC00609*	Papillary thyroid carcinoma susceptibility candidate 3
rs6759952	chr2:217406996	*DIRC3*	Disrupted in renal carcinoma 3
rs966423	chr2:217445617	*DIRC3*	Disrupted in renal carcinoma 3
rs1203952	chr20:22633494	*FOXA2*	Forkhead box A2
rs10238549	chr7:110540965	*IMMP2L*	Inner mitochondrial membrane peptidase subunit 2
rs7800391	chr7:110568186	*IMMP2L*	Inner mitochondrial membrane peptidase subunit 2
rs1799814	chr15:74720646	*CYP1A1*	Aryl hydrocarbon hydroxylase
rs7617304	chr3:158745312	*RARRES1*	Retinoic acid receptor responder 1
rs4808708	chr19:17890877	*NIS/SLC5A5*	Solute carrier family 5 member 5
rs10781500	chr9:136374886	*CARD9/SNAPC4*	Caspase recruitment domain-containing protein 9
rs1061758	chr9:34652333	*IL11RA*	Interleukin 11 receptor subunit alpha
rs10877887	chr12:62603400	*LINC01465/MIRLET7I*	Long intergenic non-protein coding RNA 1465/microRNA Let-7i

Table 3. List of the 15 SNPs associated with DTC in datasets 1 and 2.

	wPRS			PRS		
Quintile	OR(adj)	95% confidence interval	P	OR(adj)	95% confidence interval	P
I	Reference			Reference		
II	2.12	1.55 – 2.91	2.92×10^{-6}	1.43	1.04 – 1.97	0.0282
III	2.52	1.84 – 3.44	7.02×10^{-9}	2.55	1.90 – 3.40	2.87×10^{-10}
IV	3.15	2.30 – 4.32	9.65×10^{-13}	3.04	2.26 – 4.09	2.02×10^{-13}
V	6.87	4.90 – 9.64	6.12×10^{-29}	5.84	4.18 – 8.15	3.75×10^{-25}

Table 4. Odds ratio estimates for the 15-SNPs PRS quintiles. DTC state obtained in the three merged datasets was considered, using the bottom quintile (0 to 20%) as the reference group. The multivariate logistic regression model included the adjustment of ORs for age, BMI, and gender. wPRS: weighted polygenic risk score; PRS: unweighted polygenic risk score.

 Accepted Manuscript published as ETJ-22-0058.R2. Accepted for publication: 17-Aug-2022

Metric	Training set	Test set	Validation set
NPV	0.65	0.56	0.52
PPV	0.64	0.66	0.70
Sensitivity	0.88	0.87	0.85
Specificity	0.29	0.27	0.32
Accuracy	64%	64%	67%
F1-score	0.74	0.75	0.77
F0.5-score	0.67	0.70	0.73
F2-score	0.82	0.82	0.82

Notes to Table 5: NPV = Negative Predictive Value = TN/(TN+FN); PPV = Positive Predictive Value = TP/(TP+FP). F_β scores are defined as:

$$F_\beta = * \frac{(1+\beta^2) * TP}{(1+\beta^2) * TP + \beta^2 * FN + FP}$$

Table 5. Classification metrics of AdaBoost classifier on all datasets.

SUPPLEMENTARY TEXT AND FIGURES

SUPPLEMENTARY METHODS

DNA extraction

Genomic DNA was extracted from blood samples by the automated extractor EZ1 Advanced XL using the EZ1 DNA Blood Kit (Qiagen, Hilden, Germany). DNA concentrations and purity were determined by the NanoDrop™ 2000 spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA). Samples from datasets 2 and 3 were shipped in dry ice for SNP genotyping, which was performed by iPLEX® assay (Life & Brain GmbH, Bonn, Germany) as an on-demand service including primer probes design and validation.

SNP selection for association in the dataset 1

The list of 171 SNPs selected from literature was evaluated for their statistical association with the risk of DTC in the dataset 1 (Supplementary table S1). In Supplementary Table S1 significant SNPs are ranked in ascending order of the P-values of the association tests referring to the best model of inheritance (r=recessive; d=dominant; a=additive). The non-significant SNPs are reported in alphabetical order of the gene names. SNPs within the same block of linkage *disequilibrium* (LD) are reported consecutively. Within each gene, the blocks of LD are numbered progressively. For each SNP, the parameter r^2 is showed as measure of LD and it is referred to the pair-wise distance to the first listed SNP within the block. SNPs associated at nominal level of <0.05 were further genotyped in a replication set (second stage, last column), with the exception of rs1048943 because the low minor allele frequency (MAF=0.021) could not provide the adequate statistical power.

Bayesian statistics for population genetics

In the "assisted" analysis, the STRUCTURE software runs using genetic data of 15 SNPs found to be associated with DTC after the selection procedure. The case/control *status* of all individuals was

 Accepted Manuscript published as ETJ-22-0058.R2. Accepted for publication: 17-Aug-2022

known to the software and used to assist inference using the LOCPRIOR model. The Bayesian inference was launched using Markov chain Monte Carlo (MCMC) simulations after burn-in periods. The number of possible subpopulations (K) ranged from 1 to 10, for a total of 20 iterations. The most probable K number was calculated using the online tool STRUCTURE Harvester [1] applying the ΔK method [2]. Results from the 20 iterations were packaged by the Cluster Markov Packager Across K (CLUMPAK) online tool [3] using the default settings. Separate runs on unmerged datasets were also performed.

To test if this method was suitable for diverse datasets, we repeated the same protocol using dataset 3 (Table 1) as input, consisting of a balanced pool of 796 individuals, and report results using the same metrics.

Machine Learning-based disease analysis

The performance of a Machine Learning (ML) model in describing the patient/control state was evaluated. To evaluate the performance of different ML methods, an extended dataset was built by merging the two largest datasets (1 and 3), to obtain a highly informative and balanced pool of training data: randomly chosen 80% of this extended set was used to train ML models, with the case or control state known to the classifier, and the remaining 20% of the subjects was used for an initial performance evaluation, with the original labels hidden. After finding the most effective ML algorithms for the present scenario, i.e using the 15 SNPs as input variables, their predictive capability was validated on the dataset 2, to quantitatively assess their generalization capabilities. The whole procedure is summarized in the main article (Figure 1).

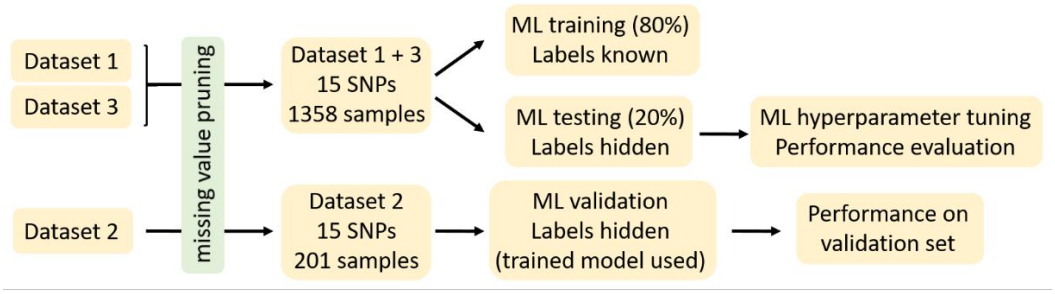
ML methodologies were applied to predict the patients' disease state based solely on the information coming from the previously selected 15 SNPs. Thus, we used the three datasets (Figure 1A) as inputs for different types of classifiers. To prepare the data and increase ML performance and robustness, any individual with missing data in the genotype was pruned from all datasets, to avoid that ML algorithms interpret missing data as a genetic information. SNP variables were encoded as follows:

firstly, each SNP genotype was split into its two alleles, thus yielding a total of 30 variables, for which ordinal variable encoding was used, to convert the four-letter genetic code (A-T-G-C) to numerical values (1-2-3-4). Since individuals with missing values were pruned from the datasets, no other numerical value was present in the training data. The final composition of the three datasets after the pruning of missing values is summarised (Table 1).

For the subsequent ML implementation, three datasets were extracted: (1) a training set, used for the training of the algorithm, composed of a randomly extracted 80% of the merged dataset 1+3; (2) a test set, for an initial performance evaluation and hyperparameter tuning, composed of the remaining 20% of the merged 1+3 dataset; (3) a validation set, corresponding to dataset 2 after pruning missing values, which constitutes a third, unseen dataset used for external validation. Indeed, it should be noted that the earlier discovery of potentially informative SNPs to genotype (see paragraph "SNP selection") did not involve any ML algorithm: since the selection was not performed based on ML performance outcome, the validation set has never been fed to the algorithm at any earlier stage.
Conversely, the rationale behind merging datasets 1 and 3 to obtain a large training set lies in the uniformity of sampling location and the goal to obtain a large pool of data for improved ML training.
A graphical summary of the ML pipeline is reported (Supplementary figure S1).
The performance of the tested ML algorithms was evaluated both on the internal test set and on the external validation set, to ensure the generalization ability of the trained algorithms and to assess the tendency towards overfitting. More in detail, the following metrics were calculated and reported: Receiver Operating Characteristic area under the curve (ROC-AUC), Accuracy, Positive Predictive Value (PPV), Negative Predictive Value (NPV), Sensitivity, Specificity, F1-Score, F0.5-score, F2-score.
In the preliminary phase of ML algorithm selection, different approaches were tested, namely k-Nearest Neighbors (kNN), Naïve Bayes (NB) [4], Random Forest (RF), Gradient Boosting (GB) [5], AdaBoost (AB) [6] and Support Vector Machine (SVM) algorithms, as implemented in the SciKit-

 Accepted Manuscript published as ETJ-22-0058.R2. Accepted for publication: 17-Aug-2022

Learn [7] library for Python. The best methodology was initially chosen based on the ROC AUC value, and subsequently further optimised for optimal performance in a hyperparameter tuning stage (Supplementary figure S1). A ROC curve describes the relationship between the sensitivity and specificity of a test by plotting the two against one another while varying the evaluation threshold, which determines the outcome of a test. Sensitivity and specificity are inversely related: as one increases the other decreases. Conventionally, since both values range between 0 and 1, the sensitivity (true positive rate) is plotted against 1 minus the specificity (false positive rate). The plot is, therefore, in essence, a representation of the trade-off between detecting true and false positive cases.



Supplementary figure S1. Pipeline of ML data analysis. Randomly chosen 80% of the merged datasets 1 and 3 served for training the ML algorithm, while the remaining, 20% was the test set. The dataset 2 was used as an external validation set. Individuals with missing SNP data were excluded from the ML analysis.

Supervised ML approaches was applied here to classify the case/control state of the individuals. Three datasets were used for the training, testing and external validation of ML algorithms, constituted respectively by 80% of merged datasets 1+3, 20% of merged datasets 1+3, and dataset 2. Among the tested algorithms, namely kNN, NB, RF, GB, AB and SVM, the most suitable was chosen based on the ROC-AUC, Sensitivity, Specificity, Accuracy and prediction probability calibration on both the test and validation sets (Table 5).

After analysing the performances of the different classifiers (Supplementary Figure S1 and S3), the Naïve Bayes classifier and the Adaptive Boosting (AdaBoost) classifier were chosen for further performance tuning and evaluation.

The Naïve Bayes classifier is a supervised learning algorithm relying on Bayes' theorem and assuming naïve conditional independence between features. In detail, in this study a Gaussian Naïve Bayes algorithm was used, as implemented in the SciKit-Learn library for Python.

The AdaBoost algorithm was implemented using decision stumps, i.e. decision trees with depth of 1, as base estimators, and the number of the latter was tuned to obtain the best AUC while at the same time optimising the computational effort. The performance of both classifiers was first evaluated on the test set, and subsequently on dataset 3 used as an external validation set, to further check the performance of classification of unseen data. Furthermore, the individual metrics including PPV, NPV, specificity, sensitivity and accuracy were analysed along with the prediction probability calibration to choose the final model, reported in the main text. Finally, the AdaBoost algorithm was found to be the most effective and well-calibrated in classifying individuals (Supplementary figure S3).

SUPPLEMENTARY RESULTS

SNPs associated with DTC

Considering that the candidate SNPs were associated with the risk of DTC in previous studies and that, in the present work, a further validation step was carried out, we considered positive the SNPs associated with the risk of DTC with a P-value below the classical statistical significance level of 0.05. Thus, 34 candidate SNPs were associated with the risk of DTC in the dataset 1 and were further evaluated in the dataset 2. The results of genotyping of this series were evaluated alone and combining the populations. Some of the SNPs were clearly associated with the risk of DTC in the dataset 2, also accomplishing the more stringent statistical significance threshold following Bonferroni's correction. Other SNPs were less clearly associated with the risk. For example, some were associated in both populations at the level of 0.05 (e.g. rs6759952), others were associated only when the two populations were combined (e.g. rs10238549). We were aware that setting a stringent threshold could

 Accepted Manuscript published as ETJ-22-0058.R2. Accepted for publication: 17-Aug-2022

lead to discharge truly positive SNPs. Thus, we relaxed the selection, by applying the criteria reported in the flowchart (Figure 2) and we drew the following conclusions:

(a) 4 SNPs (rs965513, rs3758249, rs7048394, rs944289) were robustly associated with the risk of DTC, as they accomplished the Bonferroni's threshold of the statistical significance in dataset 2;

(b) 3 SNPs (rs6759952, rs966423, and rs1203952) were considered highly likely markers of risk of DTC as they were positive in both datasets at the nominal P-value of 0.05;

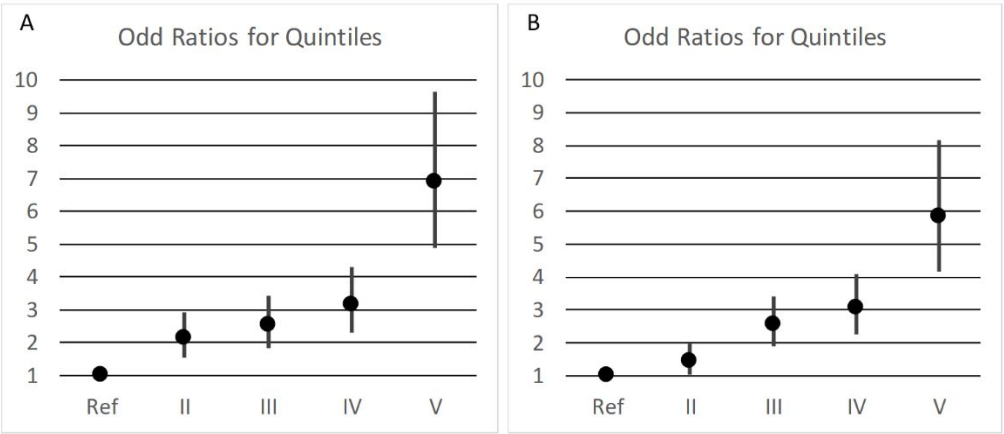
(c) 8 SNPs (rs10238549, rs7800391, rs1799814, rs7617304, rs4808708, rs10781500, rs1061758, and rs10877887) were considered as possible marker risk of DTC, as they were statistically significant at the level of 0.05 in the GWAS and in the merged dataset;

(d) the remaining 156 SNPs were considered not associated (at least in our Italian populations) with the risk of DTC: 137 were negative in the GWAS, 19 were classed as "likely negative" according to the results of the dataset 2.

Considering the number of risk alleles, and weighing each allele based on the measured OR, the ROC curve analysis demonstrated an AUC of 0.65, with 10 alleles as the cut-off to better predict the risk of having DTC.

Polygenic risk score: OR calculation

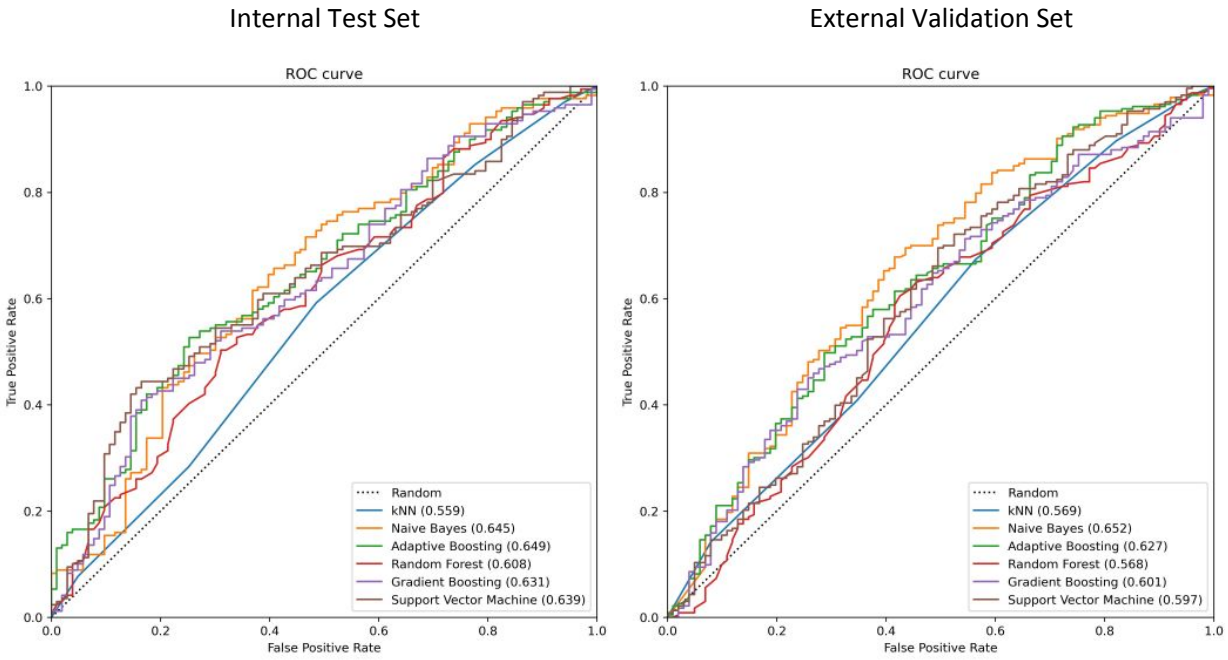
OR was calculated in each quintile under both additive and weighted models, in the three merged datasets (Table 4; Supplementary figure S2).



Supplementary figure S2. ORs calculated for each quintile. A) Multiplicative model. B) Additive model. The quintile number is indicated in the x-axis. Dots are adjusted ORs, while bars are the 95% CIs.

Choice of ML model

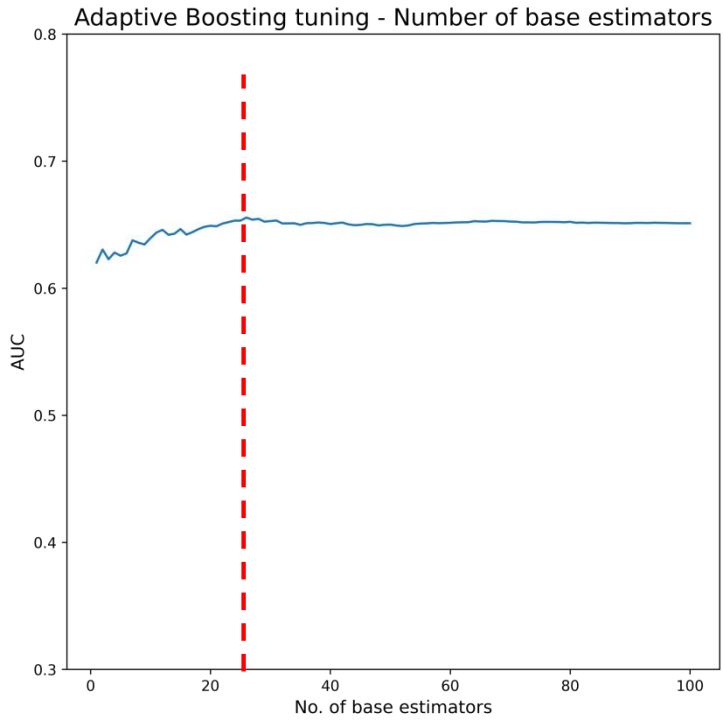
The ROC curve of different types of classifiers (kNN, NB, RF, GB, AB, SVM) on both the test and external validation sets are reported (Supplementary figure S3).



Supplementary figure S3: ROC curves of different types of classifiers. Inset reports the legend and AUC values in brackets.

Based on the initial AUC value, the AdaBoost classifier was chosen for further performance assessments and optimization. For the AdaBoost classifier, we tuned the hyperparameter regarding the number of base estimators, in a range from 1 to 100 (Supplementary figure S4).

 Accepted Manuscript published as ETJ-22-0058.R2. Accepted for publication: 17-Aug-2022



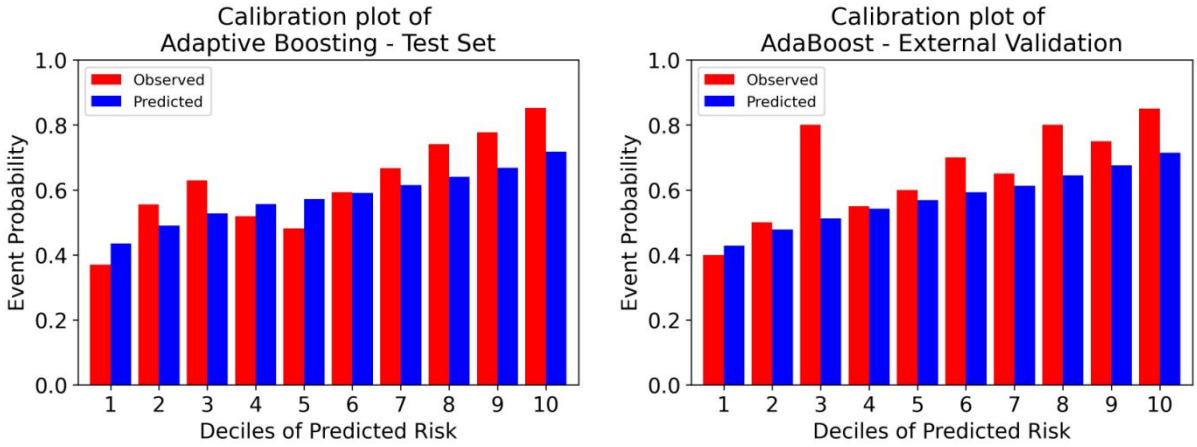
Supplementary figure S4. Tuning of number of base estimators of AdaBoost classifier based on ROC-AUC maximization. Red dashed line represents optimal value of 25.

As shown (Supplementary figure S4), we found 25 base estimators to be an optimal compromise between ROC-AUC maximization and computational cost. Further details and results regarding the optimized AdaBoost classifier are reported in the main text.

Calibration of the AdaBoost classifier

A drawback of the classifier appears related to the weak identification of the negative class (i.e. controls), as emerges from the comparably low specificity on all datasets, which has a detrimental effect also on the ROC AUC, on the negative predictive value and on overall accuracy, which however also depends on the underlying class balance of the dataset being classified. Conversely, when factoring in the class distribution (DTC vs. controls) of the underlying datasets being classified, the algorithm shows a fair confidence in predicting the DTC *status* (PPV = 0.7 on the external validation), at the cost of a comparably poorer specificity. By using Platt scaling on the predicted

probabilities, the distribution of the latter fairly matches the real distribution of DTC risk both in the test set and in the external validation set (Supplementary figure S5). This was not the case for the Naïve Bayes classifier.



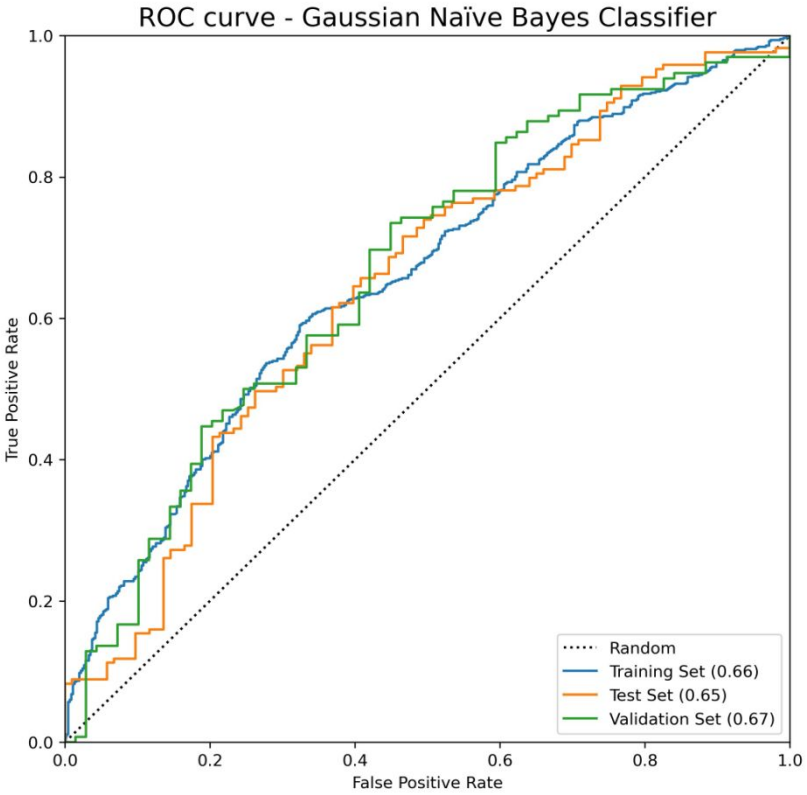
Supplementary figure S5. Calibration plot of the AdaBoost classifier on the test set (left) and the external validation set (right). Blue bars represent the mean predicted probability of DTC in each risk decile. Red bars represent the actual probability of observing DTC in each decile.

Naïve Bayes classifier

The Naïve Bayes classifier represents the second tested ML model to predict the case/control *status* using only SNP data, relying on Bayesian statistics. After training the Gaussian Naïve Bayes classifier on the dataset described in the Methods section, we deployed the classifier on the test set and obtained an AUC value of 0.65, with an overall classification accuracy of 57% and an F1 score of 0.57 (Supplementary table S4). The AUC of the precision-recall curve was 0.74, and a further 10-fold cross-validation yielded an average ROC AUC of 0.64 ± 0.02 (SD).

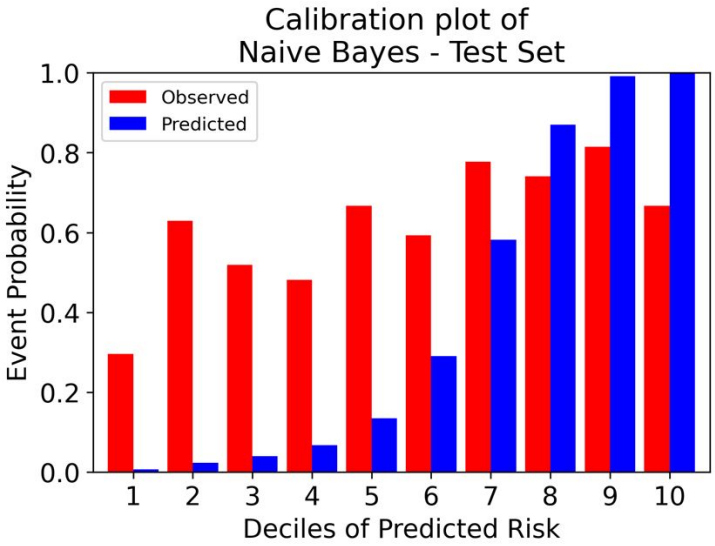
We again assessed the ability of the model to generalize to a further dataset containing unseen data by deploying the trained NB classifier on dataset 2 (Supplementary figure S6). Here, an AUC value of 0.67 was obtained, with an overall accuracy of 56% and an F1 score of 0.56.

 Accepted Manuscript published as ETJ-22-0058.R2. Accepted for publication: 17-Aug-2022



Supplementary figure S6. ROC curves obtained on all datasets. Dashed line represents random choice. A 10-fold cross-validation ROC AUC of 0.65 ± 0.03 (SD) was found.

Again, metrics from the classification of the test and validation datasets highlight no significant overfitting on the training set, with the AUC on the validation being slightly higher than the AUC obtained on both the training and test sets. In contrast to what was found for the AdaBoost classifier, this classifier appears to have a tendency towards classifying individuals into the negative class (i.e. controls), as highlighted by the higher specificity (0.76 and 0.81 on the test and validation sets, respectively). However, when factoring in the prevalence of each class in the datasets – i.e. when considering NPV and PPV values instead of specificity and sensitivity – the confidence in the positive prediction is higher (PPV is 0.76 on the test set, 0.81 on the validation set) if compared to the negative prediction (NPV values are 0.46 and 0.43 on the test and validation datasets respectively) (Supplementary figure S7).



Supplementary Figure S7. Calibration plot of the trained Naïve Bayes classifier on the test set. Blue bars represent the mean predicted probability of DTC in each risk decile. Red bars represent the actual probability of observing DTC in each decile (ground truth).

While the Naïve Bayes classifier showed comparable performance to the AdaBoost classifier in terms of raw AUC values, the calibration of the predicted probability turned out comparably poor (Supplementary figure S6). Specifically, this translates into a classifier which is overconfident in predicting the negative class, as highlighted by the skewed predicted probability towards higher deciles (Supplementary figure S7) and by the comparably high specificity and low NPV. These considerations justify the eventual choice of the AdaBoost algorithm in the light of its better probability calibration and thus increased robustness when facing imbalanced classification tasks.

SUPPLEMENTARY DISCUSSION

Three SNPs within the *FOXE1* gene (rs965513, rs3758249, rs7048394) and one in *PTCSC3-LINC00609* (rs944289) robustly associated with the risk of DTC. Moreover, other two SNPs within *DIRC3* (rs6759952 and rs966423) and one in *FOXA2* (rs1203952) were categorized as highly likely risk factors for DTC. Indeed, there is plenty of literature about the role of *FOXE1*, *PTCSC3*, *DIRC3*,

 Accepted Manuscript published as ETJ-22-0058.R2. Accepted for publication: 17-Aug-2022

and *FOXA2* in affecting the individual susceptibility to thyroid cancer [8-14], revealing they are well-established predisposing factors for DTC. Other eight SNPs, falling within seven genes and known as possible risk factors for DTC, did not replicate formally in the second stage of the study. However, they maintained or reinforced the statistical significance of the GWAS in the combined analysis. For 4 of these genes there is convincing evidence from literature for their role in thyroid tumorigenesis: SNPs within *CYP1A1* (rs1799814), *NIS-SLC5A5* (rs4808708), *IL11RA* (rs1061758), and *let-7i/LINC01465* (rs10877887) are highly likely to be true markers of individual predisposition to DTC, as well as to other cancers [15-33].

Interestingly, there is relative lack of knowledge on the role of the remaining three genes, i.e. *IMMP2L* (rs10238549, rs7800391), *RARRES1* (rs7617304), and *CARD9-SNAPC4* (rs10781500) in thyroid cancer. However, since the selected SNPs were associated with DTC in the combined analysis, they could be reasonably involved in the aetiology of the disease. *IMMP2L* encodes for the inner mitochondrial membrane peptidase subunit 2 and it is a key-molecule linking cellular senescence to metabolism and cell death signalling pathways [34]. It could be hypothesized that polymorphic variants within *IMMP2L* might affect thyrocyte ability to undergo senescence, therefore sustaining immortalization. *RARRES1*, also known as *TIG1* (tazarotene-induced gene 1), encodes for a retinoic acid regulated carboxypeptidase inhibitor. According to the function of a tumor suppressor, *RARRES1* is hypermethylated in multiple cancers [35,36] and its expression is modulated by estrogens [37], suggesting a possible role in the known gender difference of DTC incidence. Finally, *CARD9* encodes for the caspase recruitment domain-containing protein 9 and it is a central adaptor protein of innate immune responses to extracellular pathogens via releasing of active cytokines from the immune cells. In fact, suppression of T lymphocyte functioning was found in *CARD9*-knockout mice and was linked to lung cancer progression [38]. A role of *CARD9* in autoimmunity could be also evoked. Since it is involved in autoantibody-induced autoimmune disease [39], it could be hypothesized that the association with rs10781500 reflects a low-grade chronic state of autoimmunity against thyroid, predisposing carriers of the risk alleles to DTC.

The remaining 137 SNPs were not confirmed in the dataset 1, while other 19 SNPs positive in the GWAS were not confirmed in the dataset 2, at least in Italian individuals. They could have been detected as the consequence of chance findings in previously published underpowered studies, resulting as false or weakly positive signals.

A previous study evaluating PRS found 10 SNPs describing genetic predisposition to DTC for 8-11% of the total variability [40]. Among those SNPs, only three (rs2466076, rs1588635 and rs116909374) are in *linkage disequilibrium* with markers evaluated in our analysis (rs2439302, rs7028661 and rs925489, respectively) but discarded during the selection procedure. The rs1588635 falls within the *PTCSC2* gene, suggesting it as being a hot spot for genetic predisposition to DTC risk, while the others belong to genomic regions not represented by our 15-SNPs signature. Therefore, we may suppose that a number of unknown, low-penetrance SNPs contribute to DTC genetic predisposition and they may be discovered using different approaches and populations.

 Accepted Manuscript published as ETJ-22-0058.R2. Accepted for publication: 17-Aug-2022

REFERENCES

1. Earl DA, vonHoldt BM (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conservation Genetics Resources 4 (2):359-361. doi:10.1007/s12686-011-9548-7

2. Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Molecular ecology 14 (8):2611-2620. doi:10.1111/j.1365-294X.2005.02553.x

3. Kopelman NM, Mayzel J, Jakobsson M, Rosenberg NA, Mayrose I (2015) Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. Molecular ecology resources 15 (5):1179-1191. doi:10.1111/1755-0998.12387

4. Ng AY, Jordan MI (2001) On discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes. Paper presented at the Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic, Vancouver, British Columbia, Canada,

5. Friedman JH (2001) Greedy function approximation: A gradient boosting machine. The Annals of Statistics 29 (5):1189-1232, 1144

6. Freund Y, Schapire RE A desicion-theoretic generalization of on-line learning and an application to boosting. In, Berlin, Heidelberg, 1995. Computational Learning Theory. Springer Berlin Heidelberg, pp 23-37

7. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É (2011) Scikit-learn: Machine Learning in Python. J Mach Learn Res 12 (null):2825–2830

8. Gudmundsson J, Sulem P, Gudbjartsson DF, Jonasson JG, Sigurdsson A, Bergthorsson JT, He H, Blondal T, Geller F, Jakobsdottir M, Magnusdottir DN, Matthiasdottir S, Stacey SN, Skarphedinsson OB, Helgadottir H, Li W, Nagy R, Aguillo E, Faure E, Prats E, Saez B, Martinez M, Eyjolfsson GI, Bjornsdottir US, Holm H, Kristjansson K, Frigge ML, Kristvinsson H, Gulcher JR, Jonsson T, Rafnar T, Hjartarsson H, Mayordomo JI, de la Chapelle A, Hrafnkelsson J, Thorsteinsdottir U, Kong A, Stefansson K (2009) Common variants on 9q22.33 and 14q13.3 predispose to thyroid cancer in European populations. Nature genetics 41 (4):460-464. doi:10.1038/ng.339

9. Gudmundsson J, Sulem P, Gudbjartsson DF, Jonasson JG, Masson G, He H, Jonasdottir A, Sigurdsson A, Stacey SN, Johannsdottir H, Helgadottir HT, Li W, Nagy R, Ringel MD, Kloos RT, de Visser MC, Plantinga TS, den Heijer M, Aguillo E, Panadero A, Prats E, Garcia-Castaño A, De Juan A, Rivera F, Walters GB, Bjarnason H, Tryggvadottir L, Eyjolfsson GI, Bjornsdottir US, Holm H, Olafsson I, Kristjansson K, Kristvinsson H, Magnusson OT, Thorleifsson G, Gulcher JR, Kong A, Kiemeney LA, Jonsson T, Hjartarson H, Mayordomo JI, Netea-Maier RT, de la Chapelle A, Hrafnkelsson J, Thorsteinsdottir U, Rafnar T, Stefansson K (2012) Discovery of common variants associated with low TSH levels and thyroid cancer risk. Nature genetics 44 (3):319-322. doi:10.1038/ng.1046

10. Köhler A, Chen B, Gemignani F, Elisei R, Romei C, Figlioli G, Cipollini M, Cristaudo A, Bambi F, Hoffmann P, Herms S, Kalemba M, Kula D, Harris S, Broderick P, Houlston R, Pastor S, Marcos R, Velázquez A, Jarzab B,

Hemminki K, Landi S, Försti A (2013) Genome-wide association study on differentiated thyroid cancer. The Journal of clinical endocrinology and metabolism 98 (10):E1674-1681. doi:10.1210/jc.2013-1941

11. Figlioli G, Köhler A, Chen B, Elisei R, Romei C, Cipollini M, Cristaudo A, Bambi F, Paolicchi E, Hoffmann P, Herms S, Kalemba M, Kula D, Pastor S, Marcos R, Velázquez A, Jarząb B, Landi S, Hemminki K, Försti A, Gemignani F (2014) Novel genome-wide association study-based candidate loci for differentiated thyroid cancer risk. The Journal of clinical endocrinology and metabolism 99 (10):E2084-2092. doi:10.1210/jc.2014-1734

12. Son HY, Hwangbo Y, Yoo SK, Im SW, Yang SD, Kwak SJ, Park MS, Kwak SH, Cho SW, Ryu JS, Kim J, Jung YS, Kim TH, Kim SJ, Lee KE, Park DJ, Cho NH, Sung J, Seo JS, Lee EK, Park YJ, Kim JI (2017) Genome-wide association and expression quantitative trait loci studies identify multiple susceptibility loci for thyroid cancer. Nature communications 8:15966. doi:10.1038/ncomms15966

13. Wang YL, Feng SH, Guo SC, Wei WJ, Li DS, Wang Y, Wang X, Wang ZY, Ma YY, Jin L, Ji QH, Wang JC (2013) Confirmation of papillary thyroid cancer susceptibility loci identified by genome-wide association studies of chromosomes 14q13, 9q22, 2q35 and 8p12 in a Chinese population. Journal of medical genetics 50 (10):689-695. doi:10.1136/jmedgenet-2013-101687

14. Kim HS, Kim DH, Kim JY, Jeoung NH, Lee IK, Bong JG, Jung ED (2010) Microarray analysis of papillary thyroid cancers in Korean. The Korean journal of internal medicine 25 (4):399-407. doi:10.3904/kjim.2010.25.4.399

15. Figlioli G, Elisei R, Romei C, Melaiu O, Cipollini M, Bambi F, Chen B, Köhler A, Cristaudo A, Hemminki K, Gemignani F, Försti A, Landi S (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 : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology 25 (4):700-713. doi:10.1158/1055-9965.Epi-15-0652

16. Siraj AK, Ibrahim M, Al-Rasheed M, Abubaker J, Bu R, Siddiqui SU, Al-Dayel F, Al-Sanea O, Al-Nuaim A, Uddin S, Al-Kuraya K (2008) Polymorphisms of selected xenobiotic genes contribute to the development of papillary thyroid cancer susceptibility in Middle Eastern population. BMC medical genetics 9:61. doi:10.1186/1471-2350-9-61

17. Irmiakova AR, Kochetova OV, Gaĭnullina MK, Sivochalova OV, Viktorova TV (2012) [Association of polymorph variants of CYP1A2 and CYP1A1 genes with reproductive and thyroid diseases in female workers of petrochemical industry]. Meditsina truda i promyshlennaia ekologiia (5):41-48

18. Bufalo NE, Leite JL, Guilhen AC, Morari EC, Granja F, Assumpcao LV, Ward LS (2006) Smoking and susceptibility to thyroid cancer: an inverse association with CYP1A1 allelic variants. Endocrine-related cancer 13 (4):1185-1193. doi:10.1677/erc-06-0002

 Accepted Manuscript published as ETJ-22-0058.R2. Accepted for publication: 17-Aug-2022

19. GallegosVargas J, SanchezRoldan J, RonquilloSanchez M, Carmona Aparicio L, FlorianoSanchez E, CardenasRodriguez N (2016) Gene Expression of CYP1A1 and its Possible Clinical Application in Thyroid Cancer Cases. Asian Pacific journal of cancer prevention : APJCP 17 (7):3477-3482

20. de Morais RM, Sobrinho AB, de Souza Silva CM, de Oliveira JR, da Silva ICR, de Toledo Nóbrega O (2018) The Role of the NIS (SLC5A5) Gene in Papillary Thyroid Cancer: A Systematic Review. International journal of endocrinology 2018:9128754. doi:10.1155/2018/9128754

21. Heinrich PC, Behrmann I, Müller-Newen G, Schaper F, Graeve L (1998) Interleukin-6-type cytokine signalling through the gp130/Jak/STAT pathway. The Biochemical journal 334 (Pt 2) (Pt 2):297-314. doi:10.1042/bj3340297

22. Katoh M, Katoh M (2007) STAT3-induced WNT5A signaling loop in embryonic stem cells, adult normal tissues, chronic persistent inflammation, rheumatoid arthritis and cancer (Review). International journal of molecular medicine 19 (2):273-278

23. Hanavadi S, Martin TA, Watkins G, Mansel RE, Jiang WG (2006) Expression of interleukin 11 and its receptor and their prognostic value in human breast cancer. Annals of surgical oncology 13 (6):802-808. doi:10.1245/aso.2006.05.028

24. Goseki N, Koike M, Yoshida M (1992) Histopathologic characteristics of early stage esophageal carcinoma. A comparative study with gastric carcinoma. Cancer 69 (5):1088-1093. doi:10.1002/cncr.2820690503

25. Yamazumi K, Nakayama T, Kusaba T, Wen CY, Yoshizaki A, Yakata Y, Nagayasu T, Sekine I (2006) Expression of interleukin-11 and interleukin-11 receptor alpha in human colorectal adenocarcinoma; immunohistochemical analyses and correlation with clinicopathological factors. World journal of gastroenterology 12 (2):317-321. doi:10.3748/wjg.v12.i2.317

26. Vogelstein B, Fearon ER, Hamilton SR, Kern SE, Preisinger AC, Leppert M, Nakamura Y, White R, Smits AM, Bos JL (1988) Genetic alterations during colorectal-tumor development. The New England journal of medicine 319 (9):525-532. doi:10.1056/nejm198809013190901

27. Eun YG, Shin IH, Kim MJ, Chung JH, Song JY, Kwon KH (2012) Associations between promoter polymorphism -106A/G of interleukin-11 receptor alpha and papillary thyroid cancer in Korean population. Surgery 151 (2):323-329. doi:10.1016/j.surg.2011.07.014

28. Lin P, Guo YN, Shi L, Li XJ, Yang H, He Y, Li Q, Dang YW, Wei KL, Chen G (2019) Development of a prognostic index based on an immunogenomic landscape analysis of papillary thyroid cancer. Aging 11 (2):480-500. doi:10.18632/aging.101754

29. Zhong Z, Hu Z, Jiang Y, Sun R, Chen X, Chu H, Zeng M, Sun C (2016) Interleukin-11 promotes epithelial-mesenchymal transition in anaplastic thyroid carcinoma cells through PI3K/Akt/GSK3β signaling pathway activation. Oncotarget 7 (37):59652-59663. doi:10.18632/oncotarget.10831

30. Wang Y, Wei T, Xiong J, Chen P, Wang X, Zhang L, Gao L, Zhu J (2015) Association Between Genetic Polymorphisms in the Promoter Regions of Let-7 and Risk of Papillary Thyroid Carcinoma: A Case-Control Study. Medicine 94 (43):e1879. doi:10.1097/md.0000000000001879

31. Perdas E, Stawski R, Kaczka K, Zubrzycka M (2020) Analysis of Let-7 Family miRNA in Plasma as Potential Predictive Biomarkers of Diagnosis for Papillary Thyroid Cancer. Diagnostics (Basel, Switzerland) 10 (3). doi:10.3390/diagnostics10030130

32. Li M, Song Q, Li H, Lou Y, Wang L (2015) Circulating miR-25-3p and miR-451a May Be Potential Biomarkers for the Diagnosis of Papillary Thyroid Carcinoma. PloS one 10 (7):e0132403. doi:10.1371/journal.pone.0132403

33. Perdas E, Stawski R, Nowak D, Zubrzycka M (2016) The Role of miRNA in Papillary Thyroid Cancer in the Context of miRNA Let-7 Family. International journal of molecular sciences 17 (6). doi:10.3390/ijms17060909

34. Yuan L, Zhai L, Qian L, Huang D, Ding Y, Xiang H, Liu X, Thompson JW, Liu J, He YH, Chen XQ, Hu J, Kong QP, Tan M, Wang XF (2018) Switching off IMMP2L signaling drives senescence via simultaneous metabolic alteration and blockage of cell death. Cell research 28 (6):625-643. doi:10.1038/s41422-018-0043-5

35. Kloth M, Goering W, Ribarska T, Arsov C, Sorensen KD, Schulz WA (2012) The SNP rs6441224 influences transcriptional activity and prognostically relevant hypermethylation of RARRES1 in prostate cancer. International journal of cancer 131 (6):E897-904. doi:10.1002/ijc.27628

36. Yanatatsaneejit P, Chalermchai T, Kerekhanjanarong V, Shotelersuk K, Supiyaphun P, Mutirangura A, Sriuranpong V (2008) Promoter hypermethylation of CCNA1, RARRES1, and HRASLS3 in nasopharyngeal carcinoma. Oral oncology 44 (4):400-406. doi:10.1016/j.oraloncology.2007.05.008

37. Wilson CL, Sims AH, Howell A, Miller CJ, Clarke RB (2006) Effects of oestrogen on gene expression in epithelium and stroma of normal human breast tissue. Endocrine-related cancer 13 (2):617-628. doi:10.1677/erc.1.01165

38. Qu J, Liu L, Xu Q, Ren J, Xu Z, Dou H, Shen S, Hou Y, Mou Y, Wang T (2019) CARD9 prevents lung cancer development by suppressing the expansion of myeloid-derived suppressor cells and IDO production. International journal of cancer 145 (8):2225-2237. doi:10.1002/ijc.32355

39. Németh T, Futosi K, Weisinger J, Csorba K, Sitaru C, Ruland J, Mócsai A (2014) A8.25 CARD9 mediates autoantibody-induced autoimmune diseases by linking the SYK tyrosine kinase to CHEMOKINE production. Annals of the Rheumatic Diseases 73 (Suppl 1):A86-A86. doi:10.1136/annrheumdis-2013-205124.199

40. Liyanarachchi S, Gudmundsson J, Ferkingstad E, He H, Jonasson JG, Tragante V, Asselbergs FW, Xu L, Kiemeney LA, Netea-Maier RT, Mayordomo JI, Plantinga TS, Hjartarson H, Hrafnkelsson J, Sturgis EM, Brock P, Nabhan F, Thorleifsson G, Ringel MD, Stefansson K, de la Chapelle A (2020) Assessing thyroid cancer risk using polygenic risk scores. Proceedings of the National Academy of Sciences of the United States of America 117 (11):5997-6002. doi:10.1073/pnas.1919976117

 Accepted Manuscript published as ETJ-22-0058.R2. Accepted for publication: 17-Aug-2022

Supplementary table S1. List of SNPs associated to DTC in different world populations ($p<0$

Gene	SNP ID	Autors
MET	rs1621	Ning et al.
IL18R1	rs1420106	Chung et al.
IL-18	rs549908 rs360717 rs187238	Chung et al.
MIRLET7F1 promoter	rs10877887	Wang et al.
miR-34b/c	rs4938723	Chen et al.
TP53	rs1042522	Wang et al. Wu et al. Akulevich et al.
FOXE1	rs894673 rs1867277 rs3758249 rs907577 rs3021526 rs907580 rs10119760 rs704839 rs1443434	Somuncu et al. Bullock et al. Kang et al. Landa et al.
BRAF	rs3748093 rs17161747 rs1042179	Jiang et al. Zhang et al.
FAS	rs1571013 rs1800682 rs1468063	Eun et al.
FADD	rs10898853	Eun et al.
mir-149-5p	rs2292832	Wei et al.
ATM	rs373759 rs664143 rs4585 rs1801516	Song et al. Gu et al. Damiola et al.
CHEK2 *BRCA1*	rs17879961 rs16941	Wójcicka et al.
IL10	rs1800896	Çil et al. Erdogan et al.
CCL5	rs2107538	Kwon et al.
OPN	rs11730582	Mu et al.
PTCSC3/FOXE1	rs965513 rs944289 rs966423 rs2439302	Wang et al. Penna-Martinez et al. Liyanarachchi et al. Jones et al. Gudmundsson et al.
SERPINA5	rs6115 rs6112	Brenner et al.
IL22	rs2227485	Eun et al.
BIRC5	rs2071214	Wang et al.
IL1B	rs1143627 rs3136558	Ban et al.

	rs1143633 rs1143643	
TLR10	rs11466653	Kim et al.
TLR2	rs3804099 rs3804100	Kim et al.
CAPZB	rs12045440	Feng et al.
NFKB1	rs28362491	Wang et al.
XRCC3	rs861539	Wenying et al.
CASC8/CCAT2	rs6983267	Li et al.
ERBB2	rs1136201	Riaz et al. Rebaï
TERT	rs2736100	Ge et al.
MTHFR	rs1801133	Zara-Lopes et al. Yang et al.
AXIN2	rs11655966 rs3923086 rs7591	Liu et al.
HABP2	rs7080536	Bohórquez et al. Sahasrabudhe et al.
miR-146a *miR-680* *miR-933* *miR-149*	rs2910164 rs4919510 rs79402775 rs2292832	Dong et al. Jazdzewski et al.
MSH3	rs26279	Miao et al.
promoter regions of let-7	rs10877887 rs13293512	Wang et al.
FTO	rs8050136	Zhao et al.
locus 8q24	rs6983267	Sahasrabudhe et al.
sodium iodide symporter (NIS)	rs4808708 rs4808709 rs7250346	Al-Rasheed et al.
IL-18	rs360717	Abdolahi et al.
IL-27	rs153109	Zhang et al.
FOXE1	rs71369530	Pereda et al.
FOXE1	rs7028661 rs7037324 rs2997312 rs10788123 rs1254167 rs4975570	Mancikova et al. Zhu et al.
FOXE1	rs965513	Maillard et al. He et al.
TP53	rs1042522	Khan et al.
FAS	rs1571013 rs1800682 rs1468063	Eun et al.
MTHFR	rs1801133	Niu et al.
GALNTL *FOXA2*	rs7935113 rs1203952	Figlioli et al.
POU5F1B	rs6983267	Rogounovitch et al.
GNB3	rs5443	Wang and Zhang Sheu et al.

 Accepted Manuscript published as ETJ-22-0058.R2. Accepted for publication: 17-Aug-2022

IL17A *IL17RA* *IL17RB*	rs2275913 rs4819554 rs1025689	Lee et al.
ATM	rs189037 rs1800057 rs1800054 rs4986761 rs228589 rs664677	Gu et al. Xu et al. Akulevich et al.
mir-3144 *mir-608* *mir-933* *mir-449b* *mir-933*	rs67106263 rs4919510 rs10061133 rs79402775	Wei et al.
RET	rs1800861 rs1800862 rs1800863 rs3026782 rs1799939 rs1800858 rs74799832 rs77724903	Khan et al. Santos et al. Huang and Yang Lantieri et al. Elisei et al. Ceolin et al. Vaclavikova Rotondi et al.
KRAS	rs712	Jin et al.
CCND1	rs603965	Aytekin et al.
BATF *DHX35* *ARSB* *SPATA13*	rs10136427 rs7267944 rs13184587 rs1220597	Figlioli et al.
XKR4 *FOXE1*	rs2622590 rs925489	Zhan et al.
BRAF/RET	rs116909374 rs113488022 rs121913364	Guo et al. Gudmundsson et al. Fugazzola et al. Castro et al.
FOSL-2 promoter	rs925255	Kim et al.
XRCC1	rs25489 rs25487 rs1799782	Wang and Ai Hu et al. Qian et al. Ho et al.
IL1A	rs3783553	Gao et al.
NKX2-1	rs944289	Ai et al.
IL-21	rs12508721	Xiao et al.
PAX8 *STK17B*	rs4848323 rs1378624	Landa et al.
genomic region 8q24	rs10808556 rs1447295	Cipollini et al.
CCNH *ERCC5* *XPC*	rs2230641 rs2227869 rs2228001	Santos et al.

DIRC3 *IMMP2L* *IMMP2L* *RARRES1* *SNARC4/CARD9*	rs6759952 rs10238549 rs7800391 rs7617304 rs10781500	Köhler et al.
OPN	rs28357094 rs11439060 (rs11439060)	Mu et al.
ABCB1 (MDR1)	rs1045642	Ozdemir et al.
TPO	rs2048722 rs732609	Cipollini et al.
IL32	rs28372698	Plantinga et al.
ITGB2	rs2070946	Eun et al.
MDM2 *p14(ARF)*	rs2279744 rs3731217	Zhang et al.
XRCC7	rs7830743	Rahimi et al.
HRAS	rs12628	Khan et al.
RET	rs2565206	Figlioli et al.
miR-196a2	rs11614913	Wang et al.
MUTYH	rs3219489	Santos et al.
12-LOX	rs1126667	Prasad and Padma
CYP24A1	rs927650 rs2248137 rs2296241	Penna-Martinez et al.
IL1R1 promoter	rs2192752	Park et al.
IGFBP-3	rs2132571 rs2132572 rs2854744 rs13241830	Xu et al.
BIRC5	rs9904341	Yazdani et al.
locus 8q24	rs4733616	Neta et al.
OSMR	rs2278329	Hong et al.
CDH1	rs35606263	Wang et al.
BRCA1	rs799917 rs1799950	Xu et al.
CASP8 *CASP9* *BCL2*	rs3834129 rs4645978 rs2279115	Wang et al.
IL11RA	rs1061758	Eun et al.
WWOX	rs3764340	Cancemi et al.
CDK1NB	rs2066827	Pasquali et al. Hakova et al.
PTPRJ	rs4752904	Iuliano et al.
CYP1A1 *FTO* *FOXE1*	rs1799814 rs1121980 rs7048394	Figlioli et al.
ESR1	rs2228480	Rebaï et al.
GPX3	rs8177412 rs3828599 rs3805435	Lin et al.
TGFB1	rs1800472	Sigurdson et al.
P2X7	rs3751143	Dardano et al.
ADPRT (PARP1)	rs1136410	Chiang et al.
CYP1A1	rs1048943	Siraj et al.

Accepted Manuscript published as ETJ-22-0058.R2. Accepted for publication: 17-Aug-2022

/ *WDR3*	rs2145418 rs4658973	Baida et al.
BCL2 *STAT1*	rs1462129 rs10173099	Ruiz-Llorente et al.
HOX	rs920778	Zhu et al.
XRCC1	rs3213245	Halkova et al.

0.05).

Article
Journal
Int J Endocrinol. 2015;2015:405217
Exp Clin Endocrinol Diabetes. 2015 Nov;123(10):598-603.
Exp Clin Endocrinol Diabetes. 2015 Nov;123(10):598-603.
Medicine (Baltimore). 2015 Oct;94(43):e1879.
Medicine (Baltimore). 2015 Sep;94(38):e1536.
Tumour Biol. 2014 Mar;35(3):2723-8. Tumour Biol. 2014 Jan;35(1):561-5. Endocr Relat Cancer. 2009 Jun;16(2):491-503.
Int J Clin Exp Pathol. 2015 Oct 1 J Clin Endocrinol Metab. 2012 Sep;97(9):E1814-9. Tumour Biol. 2014 Jul PLoS Genet. 2009 Sep;5(9):e1000637.
Clin Endocrinol (Oxf). 2016 Mar;84(3):431-7. Thyroid. 2013 Jan;23(1):38-44.
Auris Nasus Larynx. 2015 Aug;42(4):326-31.
Eur Surg Res. 2014;52(1-2):1-7.
Int J Mol Sci. 2014 Nov 14;15(11):20968-81.
Environ Mol Mutagen. 2015 Jan Endocrine. 2014 Apr;45(3):454-61. Int J Cancer. 2014 Apr 1
Genes Chromosomes Cancer. 2014 Jun
Mol Biol Rep. 2014 May;41(5):3091-7. J Endocrinol Invest. 2008 Sep;31(9):750-4.
J Invest Surg. 2013 Dec;26(6):319-24.
Cell Physiol Biochem. 2013;32(1):171-9.
J Med Genet. 2013 Oct;50(10):689-95. Thyroid. 2014 May;24(5):845-51. Thyroid. 2013 Dec;23(12):1532-40. J Med Genet. 2012 Mar;49(3):158-63. Nat Genet. 2012 Jan 22;44(3):319-22.
PLoS One. 2013;8(3):e57243.
J Endocrinol Invest. 2013 Sep;36(8):584-7.
Pathol Res Pract. 2013 Mar;209(3):151-4.
Immunol Invest. 2012;41(8):888-905.

Accepted Manuscript published as ETJ-22-0058.R2. Accepted for publication: 17-Aug-2022

Endocrine. 2013 Feb;43(1):161-9.

J Korean Med Sci. 2012 Nov;27(11):1333-8.

Int J Endocrinol. 2015;2015:250542.

Genet Test Mol Biomarkers. 2015 Mar;19(3):167-71.

Med Sci Monit. 2015; 21: 3978–3985.

Med Sci Monit. 2016 Jun 2;22:1866-71.

Arch Iran Med. 2016 Jun;19(6):430-8

Genet Test Mol Biomarkers. 2009 Dec;13(6):779-84.

Sci Rep. 2016 May 17;6:26037.

Genet Mol Res. 2016 May 9;15(2).

Eur Rev Med Pharmacol Sci. 2014;18(15):2097-101.

Iran Red Crescent Med J. 2016 Jan 1;18(2):e20960

Endocr Connect. 2016 May;5(3):123-7.

J Clin Endocrinol Metab. 2015 Dec 21:jc20153928.

Int J Clin Exp Pathol. 2015; 8(10): 13450–13457.

Proc Natl Acad Sci U S A. 2008 May 20;105(20):7269-74.

Int J Clin Exp Pathol. 2015 Sep 1;8(9):11060-7.

Medicine (Baltimore). 2015 Oct;94(43):e1879.

Fam Cancer. 2016 Jan;15(1):145-53.

Endocr Relat Cancer. 2015 Oct;22(5):841-9.

Gene. 2015 Nov 10;572(2):163-8.

Endocrine. 2015 Dec;50(3):698-707

Tumour Biol. 2015 Sep;36(10):8207-11.

BMC Genet. 2015 Mar 1;16:22. doi: 10.1186/s12863-015-0180-5.

Int J Cancer. 2015 Oct 15;137(8):1870-8.

PLoS One. 2014 Jan 29;9(1):e87332.

PLoS One. 2015 Apr 7;10(4):e0123700

J Clin Endocrinol Metab. 2015 Jan;100(1):E164-72.

Cancer Biomark. 2015;15(4):459-65.

Auris Nasus Larynx. 2015 Aug;42(4):326-31.

Dis Markers. 2015;2015:681313.

Sci Rep. 2015 Mar 10;5:8922.

Thyroid. 2015 Mar;25(3):333-40.

J BUON. 2014 Oct-Dec;19(4):1092-5.

J Pathol. 2007 Jan;211(1):60-6.

Cytokine. 2015 Feb;71(2):283-8.

Int J Endocrinol. 2014;2014:370825.
J Clin Endocrinol Metab. 2012 Jun;97(6):1913-21.
Endocr Relat Cancer. 2009 Jun;16(2):491-503.

Endocrine. 2015 Jun;49(2):436-44.

J Cell Biochem. 2015 Aug;116(8):1712-8.
PLoS One. 2014 Oct 17;9(10):e109822.
Int J Clin Exp Pathol. 2015 May 1;8(5):5793-7.
Int J Cancer. 2013 Jun 15;132(12):2808-19.
J Clin Endocrinol Metab. 2004 Jul;89(7):3579-84.
Int J Mol Sci. 2012;13(1):221-39.
Endocrine. 2009 Dec;36(3):419-24.
J Endocrinol Invest. 2009 Feb;32(2):115-8.

Med Oncol. 2014 Oct;31(10):221.

Asian Pac J Cancer Prev. 2014

J Clin Endocrinol Metab. 2014 Oct;99(10):E2084-92.

Hum Mol Genet. 2014 Oct 15;23(20):5505-17.

Cancer Med. 2014 Jun;3(3):731-5.
Nat Genet. 2012 Jan 22;44(3):319-22.
Endocr Relat Cancer. 2006 Jun;13(2):455-64.
J Clin Endocrinol Metab. 2006 Jan;91(1):213-20.

Clin Exp Otorhinolaryngol. 2014 Mar;7(1):42-6.

Tumour Biol. 2014 May;35(5):4791-7.
Gene. 2013 Oct 10;528(2):67-73.
Asian Pac J Cancer Prev. 2012;13(12):6385-90.
Thyroid. 2009 Feb;19(2):129-35.

Tumour Biol. 2014 Apr;35(4):3861-5.

Front Med. 2014 Mar;8(1):113-7.

Gene. 2014 Mar 1;537(1):15-9.

PLoS One. 2013 Sep 23;8(9):e74765.

Cancer Epidemiol Biomarkers Prev. 2013 Nov;22(11):2121-5.

Oncol Rep. 2013 Nov;30(5):2458-66.

Accepted Manuscript published as ETJ-22-0058.R2. Accepted for publication: 17-Aug-2022

J Clin Endocrinol Metab. 2013 Oct;98(10):E1674-81.
Cell Physiol Biochem. 2013;32(1):171-9.
Asian Pac J Cancer Prev. 2013;14(5):3213-7.
Int J Cancer. 2013 Dec 15;133(12):2843-51.
Carcinogenesis. 2013 Jul;34(7):1529-35.
Am J Surg. 2013 Jun;205(6):631-5.
Surgery. 2013 May;153(5):711-7.
Iran Biomed J. 2012;16(4):218-22.
Tumour Biol. 2013 Feb;34(1):521-9.
Mutat Res. 2013 Jan-Mar;752(1):36-44.
Mutagenesis. 2012 Nov;27(6):779-88.
Oncol Rep. 2012 Nov;28(5):1859-68.
Fam Cancer. 2012 Dec;11(4):615-21.
Thyroid. 2012 Jul;22(7):709-16.
Int J Immunogenet. 2012 Dec;39(6):501-7.
Mol Carcinog. 2012 Oct;51 Suppl 1:E158-67.
Pathol Res Pract. 2012 Feb 15;208(2):100-3.
Laryngoscope. 2012 May;122(5):1040-2.
Clin Exp Otorhinolaryngol. 2011 Dec;4(4):193-8.
Endocrine. 2012 Jun;41(3):526-31.
Thyroid. 2012 Jan;22(1):35-43.
Med Oncol. 2012 Dec;29(4):2445-51.
Surgery. 2012 Feb;151(2):323-9.
Int J Cancer. 2011 Dec 15;129(12):2816-24.
Eur J Endocrinol. 2011 Mar;164(3):397-404.
Cancer Biomark. 2016 Jun 7;17(1):97-106.
Endocr Relat Cancer. 2010 Oct 29;17(4):1001-6.
Cancer Epidemiol Biomarkers Prev. 2016 Apr;25(4):700-13.
J Recept Signal Transduct Res. 2009;29(2):113-8.
Surgery. 2009 May;145(5):508-13.
Radiat Res. 2009 Jan;171(1):77-88.
J Clin Endocrinol Metab. 2009 Feb;94(2):695-8.
Clin Cancer Res. 2008 Sep 15;14(18):5919-24.
BMC Med Genet. 2008 Jul 5;9:61.

Cancer Epidemiol Biomarkers Prev. 2008 Jun;17(6):1499-504.
Cancer Res. 2007 Oct 1;67(19):9561-7.
Sci Rep. 2016 Aug 23;6:31969.
Cancer Biomark. 2016 Jun 7;17(1):97-106.

Accepted Manuscript published as ETJ-22-0058.R2. Accepted for publication: 17-Aug-2022

Population	**Geographic area**
Chinese	China
Korean	Korea
Korean	Korea
Chinese	China
Chinese	China
Chinese Meta-analysis (several ethnic groups) Caucasians	China Various Ukraine
Wellcome Trust Case Control Consortium samples Spanish	China UK Meta-analysis Spain
Chinese Chinese	China China
Korean	Korea
Korean	Korea
	China
korean	Korea Belarus
Polish	Poland
Turkish Turkish	Turkey Turkey
Korea	Korea
Chinese	China
Chinese German-Caucasian American, Polish Several ethnic groups Icelandic, Dutch, Americans, Spanish	China Germany Ohio (USA), Poland Various Iceland, Holland, USA, Spain
	Korea
China	China
Korean	Korea

Korean	Korea
Korean	Korea
Chinese	China
Chinese	China
Mata-analysis (several ethnic groups)	Various
Mata-analysis (several ethnic groups)	Various
African-American, Caucasian Tunisian	Various Tunisia
Various	Various
Various Chinese	Various China
Chinese	China
Caucasian (but not Hispanics)	British Isles
Han Chinese Finnish, Polish, Americans	China Finland, Poland, USA
Mata-analysis (several ethnic groups)	Various
Asian	
Meta-analysis (several ethnic groups)	Various
British Isles, Colombia and Japan	
Iranian (protective effect from PTC)	Iran
Chinese	China
Cuban (confirmed in Japanese and European)	Cuba
Southern European Various	Various
French Polynesian population Chinese	Polinesia China
Kashmir Valley people	Kashmir Valley
Mata-analysis (several ethnic groups)	Various
Italians	Italy
Japanese	Japan
Mata-analysis (several ethnic groups) German	Various Germany

 Accepted Manuscript published as ETJ-22-0058.R2. Accepted for publication: 17-Aug-2022

Korean	Korea
Chinese Several ethnic groups Caucasians	Northern China Brazil, Texas (USA) Ukraine
Chinese	China
Northern Indian Portuguese Chinese Meta-analysis (several ethnic groups) Italians Review paper Czech Italians	North India Portugal China Various Italy Review paper Czech Republic Italy
Chinese	China
Turkish	Turkey
Italian, Polish, Spanish	Italy, Poland, Spain
Chinese	China
Various Icelandic, Dutch, Americans, Spanish Italians	Various Iceland, Holland, USA, Spain Italy
Korean	Korea
Meta-analysis (several ethnic groups) Meta-analysis (several ethnic groups) Meta-analysis (several ethnic groups)	Various Various Various
Chinese	USA China
Han Chinese	Northern China
Chinese	China
Spanish	Spain
Central-Southern Italian	Italy
Portuguese-Caucasian	Portugal

Italian	Italy
Chinese	China
Italian, Spanish	Italy, Spain
Korean	Korea
Mixed ethnic groups	USA
Iranian	Iran
Kashmiri	Kashmir Valley
Meta-analysis (several ethnic groups)	Various
Meta-analysis (several ethnic groups)	Various
Portuguese	Portugal
German	Germany
Korean	Korea
Caucasian Americans	Texas (USA)
Iranian	Iran
Several ethnic groups	USA
Korean	Korea
Han Chinese	China
Several ethnic groups	USA
Han Chinese	China
Korean	Korea
Italians	Italy
Italians Caucasians	Italy Russia
Finnish, French, Italians (Caucasians)	Europe
Meta-analysis (several ethnic groups)	Various
Tunisian	Tunisia
Chinese	China
Kazakh or Russian	Kazakhstan
Italians	Italy
Chinese	China
Saudi Arabian	Saudi Arabia

 Accepted Manuscript published as ETJ-22-0058.R2. Accepted for publication: 17-Aug-2022

Spanish	Spain
Spanish	Spain
Shandong, Jiangsu and Jilin Chinese	China
Caucasians	Russia

Supplementary table 2. List of 171 SNPs selected from literature and evaluated for their statistical associat
1000Genomes database (www.internationalgenome.org); Best P-value (model) a=additive, when the mode

Gene region	Queried SNPs	A1	A2	Chr	r2	LD block
PTCSC2-FOXE1	rs965513g	G	A	9		1
	rs925489i	T	C	9	0.98	
	rs7028661i	G	A	9	0.94	
	rs3758249i	A	T	9		2
	rs12348691i	A	G	9	1	
	rs1443435g	G	A	9	1	
	rs3021526i	T	C	9	1	
	rs1443434i	T	G	9	1	
	rs907577i	A	G	9	0.99	
	rs1867277i	G	A	9	0.94	
	rs10119760i	C	G	9	0.83	
	rs925487g	A	G	9	0.80	
	rs7048394i	C	T	9		3
	rs907580i	G	A	9	0.99	
	rs10759960g	A	G	9	0.99	
	rs7037324i	G	A	9	0.79	
IMMP2L (LOC105375451)	rs10238549g	C	T	7		1
	rs7800391g	C	T	7		2
ARSB	rs13184587g	G	A	5		1
BATF (LOC105370572)	rs10136427g	G	A	14		1
DIRC3	rs6759952g	T	C	2		1
	rs966423g	C	T	2		2
DHX35-LINC01734	rs7267944g	T	C	20		1
SPATA13	rs1220597g	A	G	13		1
FOXA2-LINC01384-LINC01747	rs1203952g	A	G	20		1
GALNT18	rs7935113g	A	G	11		1
CYP1A1	rs1799814g	C	A	15		1
	rs1048943i	T	C	15		2
TLR10	rs11466653g	T	C	4		1
FTO	rs1121980g	C	T	16		1
	rs8050136g	C	A	16	0.88	
RARRES1 (LOC100287290)	rs7617304g	G	A	3		
NIS- SLC5A5	rs35036312i	G	A	19		1
	rs12327843g	T	C	19	0.99	
	rs4808709i	A	G	19	0.92	
	rs4808708g	G	A	19	0.86	
CARD9-SNAPC4	rs10781500g	C	T	9		1
WDR-SPAG17	rs4658973i	T	G	1		1
	rs1321665g	A	C	1	0.87	
CDKN1B	rs2066827g	A	C	12		1
MET	rs1621g	A	G	7		1
IL10	rs1800896i	T	C	1		1
	rs3024502g	C	T	1	0.98	
AXIN2	rs11655966i	A	T	17		1
	rs7591g	A	T	17		2

 Accepted Manuscript published as ETJ-22-0058.R2. Accepted for publication: 17-Aug-2022

	rs3923086g	A	C	17		3
PTCSC3-LINC00609	rs944289g	G	A	14		1
STAT1	rs10173099i	C	T	2		1
	rs2030171g	G	A	2	1	
IL17RB	rs1025689i	C	G	3		1
	rs4687751g	A	G	3	0.99	
IL11RA	rs1061758g	G	A	9		1
HABP2	rs7080536i	G	A	10		1
ESR1	rs2228480g	G	A	6		1
ALOX12	rs1126667g	G	A	17		1
LINC01465	rs10877887i	T	C	12		1
	rs11174538g	A	C	12	0.94	
	rs12357g	G	A	12		2
ABCB1	rs1045642g	G	A	7		1
CASC11 (AC104370)	rs4733616g	C	T	8		1
ATF2	rs79402775i	G	A	2		1
ATM	rs1800054i	C	G	11		1
	rs1800057i	C	G	11		2
	rs1801516i	G	A	11		3
	rs189037i	G	A	11		4
	rs228589i	T	A	11	0.99	
	rs664677i	T	C	11	0.89	
	rs609429i	G	C	11	0.80	
	rs609429i	G	C	11		5
	rs664143g	G	A	11	0.95	
	rs4585g	C	A	11	0.93	
	rs373759i	C	T	11		6
	rs4986761i	T	C	11		7
BCL2	rs1462129i	T	C	18		1
	rs2279115i	A	C	18	0.98	
BIRC5	rs2071214i	A	G	17		1
	rs9904341i	G	C	17		2
BRAF	rs17161747i	G	C	7		1
BRCA1	rs16941i	T	C	17		1
	rs799917g	G	A	17	0.90	
	rs1799950g	A	G	17		2
BTG4	rs4938723i	T	C	11		1
CAPZB	rs12045440i	T	G	1		1
CASC8	rs6983267g	C	A	8		1
	rs3834129i		delAGTA	2		2
CASP9	rs4645978i	G	A	1		1
CCL5	rs2107538i	C	T	1		1
CCND1	rs603965g	G	A	11		1
CCNH	rs2230641i	T	C	5		1
CDC20B	rs10061133i	A	G	5		1
CDKN2A	rs3731217g	A	C	9		1
CYP24A1	rs2248137i	G	C	20		1
	rs2296241g	G	A	20		2
	rs927650g	G	A	20		3

ERCC5	rs2227869i	G	C	13		1
FADD	rs10898853i	T	C	11		1
FAS	rs1468063g	A	G	10		1
	rs1800682g	G	A	10		2
	rs1571013i	A	G	10		3
FOSL2	rs925255g	A	G	2		1
GNB3	rs5443g	A	G	12		1
GPC1	rs2292832g	A	G	2		1
GPX3	rs3805435g	G	A	5		1
	rs3828599g	A	G	5		2
	rs8177412g	G	A	5		3
HOTAIR	rs920778i	G	A	12		1
HOXC5	rs11614913g	A	G	12		1
IGFBP3	rs13241830i	G	A	7		1
	rs2132571i	T	C	7	0.97	
	rs2132572i	C	T	7		2
	rs2854744i	G	T	7		3
IL17A	rs2275913i	G	A	6		1
IL17RA	rs4819554g	G	A	22		1
IL18	rs187238i	C	G	11		1
	rs360717g	A	G	11	1	
	rs549908g	C	A	11	0.84	
IL18RAP	rs1420106g	A	G	2		1
IL1A	rs3783553i		dupTGAA	2		1
IL1B	rs1143633g	A	G	2		1
	rs1143643i	C	T	2	0.99	
	rs3136558g	G	A	2		2
	rs1143627g	G	A	2		3
IL1R1	rs2192752i	G	T	2		1
IL21	rs12508721i	C	T	4		1
IL22	rs2227485g	A	G	12		1
IL27	rs153109g	G	A	16		1
IL32	rs28372698i	T	A	16		1
ITGB2	rs2070946g	G	A	21		1
KLC1	rs861539g	A	G	14		1
KRAS	rs712i	A	C	12		1
LOC101928570	rs4075570i	A	G	6		1
LRRC56	rs12628i	A	G	11		1
MBIP	rs116909374i	C	7	14		1
MDM2	rs2279744i	T	G	12		1
MIR3144	rs67106263i	G	A	6		1
MIRLET7A1	rs13293512i	G	A	6		1
MSH3	rs26279g	A	G	5		1
MTHFR	rs1801133g	G	A	1		1
MUTYH	rs3219489i	C	G	1		1
NFKB1	rs28362491i		delATTG	4		1
NRG1	rs2439302i	C	G	8		1
OSMR	rs2278329i	G	A	5		1
P2RX7	rs3751143g	A	C	12		1
PARP1	rs1136410g	A	G	1		1

 Accepted Manuscript published as ETJ-22-0058.R2. Accepted for publication: 17-Aug-2022

PAX8	rs4848323g	G	A	2		1
PRKDC	rs7830743g	A	G	8		1
PTPRJ	rs4752904g	C	G	11		1
RET	rs1799939g	G	A	10		1
	rs1800863i	C	G	10	0.97	
	rs3026782i	G	A	10	0.96	
	rs1800858i	G	A	10		2
	rs1800861i	A	C	10		3
	rs1800862i	C	T	10		4
	rs2565206i	C	A	10		5
	rs77724903i	A	T	10		6
RP11-323P17.1	rs10788123g	G	A	10		1
RP11-382A18.1	rs10808556g	A	G	8		2
	rs1447295g	C	A	8		3
SEMA4G	rs4919510g	C	G	10		1
SERPINA5	rs6112g	G	A	14		1
	rs6115i	A	G	14	0.94	
SPP1	rs11439060i		dupG	4		1
	rs11730582g	A	G	4		2
	rs28357094i	T	G	4		3
STK17B	rs1378624i	G	A	4		1
TERT	rs2736100g	A	C	5		1
TGFB1	rs1800472g	G	A	19		1
TLR2	rs3804099g	A	G	4		1
	rs3804100g	A	G	4		2
TP53	rs1042522g	G	C	17		1
TPO	rs2048722g	G	A	2		1
	rs732609g	A	C	2	0.80	
WDR11-AS1	rs1254167i	C	G	10		1
	rs2997312g	G	A	10		1
WDR3-TBX15	rs2145418i	G	A	10		1
WWOX	rs3764340i	C	G	16		1
XKR4	rs2622590i	G	A	8		1
XPC	rs2228001g	A	C	3		1
XRCC1	rs1799782g	G	A	19		1
	rs25487i	C	T	19		2
	rs25489g	G	A	19		3
	rs3213245i	A	G	19		4

tion with the risk of DTC in the dataset 1. Querid SNPs g= directly genotyped; i= imputed from
el tests whether heterozygotes have intermediate risk between common and rare homozygotes;

Controls			**Cases**			**Best P-value (model)**	**Second stage**
+/+	+/-	-/-	+/+	+/-	-/-		
199	185	47	191	319	139	5.05x10-10 (a)	Yes
197	187	47	189	321	139		
197	187	47	189	321	139		
139	226	65	137	325	187	1.01x10-8 (a)	Yes
139	226	65	137	325	187		
139	226	65	137	325	187		
139	226	65	137	325	187		
139	226	65	137	325	187		
139	226	65	137	325	187		
139	226	65	137	325	187		
162	215	53	168	319	162		
162	215	53	168	319	162		
206	192	33	236	306	105	2.30x10-6 (a)	Yes
206	192	33	236	306	105		
206	192	33	236	306	105		
158	219	54	165	320	164		
169	207	55	335	261	53	3.16x10-5 (a)	Yes
183	201	46	220	303	126	6.22x10-5 (a)	Yes
208	183	40	386	233	30	3.36x10-5 (a)	Yes
286	128	17	497	142	10	6.75x10-5 (a)	Yes
166	202	63	196	297	156	7.61x10-5 (a)	Yes
151	219	61	208	295	146	6.47x10-4 (r)	Yes
285	133	13	361	241	47	7.69x10-5 (a)	Yes
160	205	66	187	304	158	1.02x10-4 (a)	Yes
267	147	15	333	267	47	1.12x10-4 (a)	Yes
306	110	15	386	230	31	1.43x10-4 (d)	Yes
389	41	1	541	101	7	8.57x10-4 (a)	Yes
394	17	0	576	21	0	0.0176 (d)	No
402	28	0	567	80	2	8.83x10-4 (a)	Yes
214	182	34	320	280	47	1.15x10-3 (a)	Yes
130	205	96	234	313	102		Yes
241	165	24	305	284	60	1.23x10-3 (a)	Yes
265	150	15	351	251	47	2.63x10-3 (a)	Yes
265	150	15	351	251	47		
264	149	18	348	248	53		
264	149	18	348	248	53		Yes
164	196	71	297	278	74	2.85x10-3 (a)	Yes
116	229	86	229	294	126	3.86x10-3 (d)	Yes
116	227	88	229	293	127		
216	188	27	315	242	68	9.99x10-3 (r)	Yes
169	215	46	265	280	104	0.0133 (r)	Yes
176	189	66	262	320	67	0.0145 (r)	Yes
176	189	66	262	320	67		
236	110	9	337	166	32	0.0163 (r)	Yes
151	216	64	234	295	118	0.140 (r)	

Accepted Manuscript published as ETJ-22-0058.R2. Accepted for publication: 17-Aug-2022

97	232	102	172	313	164	0.130 (d)	
142	200	89	259	280	110	0.0174 (a)	Yes
158	195	49	218	329	102	0.031 (a)	Yes
158	195	49	218	329	102		
138	209	84	247	298	104	0.0301 (a)	Yes
138	209	84	247	298	104		
313	108	10	436	187	26	0.0350 (a)	Yes
405	22	0	626	18	0	0.0460 (d)	Yes
308	112	11	495	145	9	0.0480 (a)	Yes
164	188	79	220	338	90	0.0493 (r)	Yes
167	207	55	222	309	112	0.0363 (a)	Yes
167	207	55	222	309	112		
202	185	44	278	289	82	0.123 (a)	
105	207	119	187	302	160	0.100 (a)	
286	130	15	434	191	24	0.850 (a)	
406	19	0	606	32	1	0.600 (d)	
414	2	0	629	3	0	0.988 (d)	
400	12	0	595	19	0	0.867 (d)	
307	112	7	484	140	14	0.165 (d)	
217	180	34	309	279	61	0.289 (a)	
223	177	31	318	271	60	0.234 (r)	
223	177	31	318	271	60	0.229 (a)	
216	179	34	309	278	60	0.326 (a)	
216	179	34	309	278	60	0.326 (a)	
225	175	31	319	272	58	0.227 (a)	
216	179	35	308	271	58	0.479 (a)	
136	207	88	212	316	121	0.470 (r)	
414	6	0	623	10	0	0.844 (d)	
133	210	88	202	321	126	0.685 (r)	
133	210	88	202	321	126	0.685 (r)	
391	33	1	603	36	1	0.158 (a)	
189	186	56	301	287	61	0.063 (r)	
382	21	1	569	32	1	0.776 (r)	
206	181	43	306	275	64	0.881 (d)	
198	185	48	291	282	76	0.689 (a)	
344	81	6	541	104	3	0.720 (a)	
157	198	76	254	299	94	0.171 (r)	
178	199	54	305	278	66	0.052 (a)	
117	212	102	163	339	147	0.456 (d)	
129	225	77	211	294	144	0.085 (r)	
134	196	101	172	325	152	0.101 (d)	
294	128	8	443	182	24	0.082 (r)	
109	208	114	162	331	156	0.370 (r)	
234	176	21	374	237	38	0.279 (d)	
331	94	5	490	150	9	0.555 (a)	
318	104	9	484	151	14	0.770 (d)	
141	201	89	177	317	155	0.052 (a)	
125	203	103	197	305	147	0.560 (a)	
112	223	96	185	310	154	0.360 (d)	

377	51	1	586	61	1	0.182 (a)
202	186	43	289	281	79	0.264 (r)
356	69	6	508	133	7	0.090 (d)
128	216	87	181	339	129	0.510 (d)
149	206	76	219	319	111	0.778 (d)
175	200	56	254	318	77	0.580 (r)
190	191	50	273	301	75	0.512 (d)
218	177	36	325	279	45	0.386 (r)
344	84	3	497	138	13	0.120 (r)
228	171	32	315	279	55	0.160 (r)
319	101	11	479	156	14	0.673 (r)
149	213	69	250	311	88	0.131 (a)
176	210	45	292	267	71	0.076 (d)
207	190	34	307	272	70	0.114 (r)
207	190	34	307	272	70	0.114 (r)
305	118	8	478	155	15	0.279 (d)
124	227	80	192	331	126	0.727 (r)
187	161	43	273	258	51	0.247 (r)
281	137	13	434	194	21	0.569 (d)
235	171	25	382	232	34	0.150 (d)
235	171	25	382	232	34	0.150 (d)
221	174	35	354	256	39	0.172 (a)
265	149	17	398	222	27	0.853 (r)
213	183	30	330	364	49	0.065 (d)
188	186	57	272	296	79	0.608 (d)
188	186	57	272	296	79	0.608 (d)
270	148	13	414	202	33	0.099 (r)
173	209	49	261	299	88	0.285 (r)
201	180	50	310	274	65	0.408 (r)
178	199	54	296	289	64	0.091 (a)
114	223	94	185	337	127	0.316 (a)
175	205	50	271	304	74	0.730 (d)
68	78	22	94	110	26	0.588 (r)
242	155	33	366	237	41	0.407 (r)
146	208	77	199	328	122	0.268 (d)
136	220	75	187	326	136	0.145 (a)
187	183	61	277	294	78	0.305 (r)
272	140	19	420	204	22	0.399 (r)
388	15	1	553	29	1	0.385 (d)
173	196	62	254	273	108	0.251 (r)
266	148	17	418	210	21	0.327 (a)
216	187	28	350	253	45	0.209 (d)
179	205	47	288	288	73	0.355 (d)
131	203	96	173	321	154	0.178 (d)
263	146	22	398	216	35	0.835 (r)
190	200	36	300	279	67	0.296 (r)
135	219	77	185	329	135	0.183 (a)
429	2	0	647	2	0	0.700 (a)
256	157	18	370	241	38	0.223 (r)
314	108	9	480	158	10	0.505 (r)

Accepted Manuscript published as ETJ-22-0058.R2. Accepted for publication: 17-Aug-2022

151	209	71	256	294	99	0.143 (d)
361	67	3	537	107	5	0.660 (a)
121	233	77	180	333	136	0.211 (r)
277	133	21	384	237	27	0.098 (d)
277	133	21	384	237	27	0.098 (d)
277	133	21	384	237	27	0.098 (d)
238	162	31	385	229	35	0.119 (a)
248	158	25	381	231	37	0.704 (d)
379	49	2	573	67	6	0.385 (r)
199	193	39	302	277	69	0.391 (r)
435	0	0	645	0	0	-
230	168	33	329	273	47	0.390 (d)
178	186	67	236	332	80	0.106 (d)
376	51	4	560	89	0	0.991 (a)
280	136	15	425	202	22	0.859 (a)
210	174	46	307	276	66	0.622 (d)
210	174	46	307	276	66	0.622 (d)
184	200	42	306	270	67	0.158 (d)
104	232	95	172	307	170	0.120 (r)
249	160	17	402	217	26	0.204 (d)
211	168	36	326	239	53	0.407 (r)
131	215	85	213	327	109	0.213 (a)
377	54	0	560	84	5	0.378 (a)
140	205	86	202	320	127	0.639 (d)
379	51	1	555	94	0	0.309 (a)
239	157	34	366	243	40	0.267 (r)
136	214	79	206	329	114	0.722 (r)
141	219	71	226	325	97	0.586 (d)
296	130	5	465	172	10	0.260 (d)
296	130	5	465	172	10	0.260 (d)
271	146	14	419	205	25	0.573 (d)
387	41	3	567	79	2	0.249 (d)
348	74	9	496	148	5	0.061 (r)
123	217	91	176	309	150	0.337 (r)
364	64	2	545	98	6	0.640 (a)
186	200	45	317	262	70	0.066 (d)
359	69	3	556	91	2	0.238 (a)
160	206	64	233	300	116	0.197 (r)

Table 2. Association analyses for 34 SNPs genotyped in the dataset 2 set alone (r) and combined (c) SNPs are listed as reported in table 1. Odd Ratios (adjusted for sex, age, smoking habit, body mass i significant at nominal level of 0.05 and evaluated for the Bonferroni's correction (p-threshold=1.47x

		Heterozygote		Homozygote
PTCSC2-FOXE1		OR_{adj} (95% CI)	P_{ass}	OR_{adj} (95% CI)
rs965513	r	1.27 (0.84-1.91)	0.25	2.57 (1.48-4.47)
	c	1.76 (1.41-2.20)	$8.35x10^{-7}$	3.23 (2.35-4.43)
rs3758249	r	1.51 (0.99-2.29)	0.056	2.05 (1.27-3.31)
	c	1.62 (1.27-2.06)	$8.80x10^{-5}$	2.62 (1.95-3.50)
rs7048394	r	1.31 (0.90-1.91)	0.15	1.81 (0.96-3.41)
	c	1.45 (1.17-1.79)	$7.06x10^{-4}$	2.92 (1.99-4.29)
IMMP2L				
rs10238549	r	1.07 (0.74-1.55)	0.73	0.78 (0.41-1.46)
	c	0.80 (0.65-0.99)	0.044	0.53 (0.37-0.76)
rs7800391	r	1.06 (0.73-1.55)	0.76	1.13 (0.67-1.91)
	c	1.23 (0.99-1.53)	0.064	1.72 (1.27-2.33)
ARSB				
rs13184587	r	0.86 (0.58-1.27)	0.45	1.75 (0.89-3.47)
	c	0.76 (0.61-0.94)	0.013	0.69 (0.46-1.05)
BATF				
rs10136427	r	0.71 (0.47-1.06)	0.092	1.5 (0.34-6.61)
	c	0.70 (0.56-0.88)	$2.63x10^{-3}$	0.71 (0.35-1.47)
DIRC3				
rs6759952	r	0.97 (0.65-1.45)	0.88	1.77 (1.09-2.86)
	c	1.09 (0.87-1.37)	0.47	1.69 (1.27-2.25)
rs966423	r	1.04 (0.69-1.56)	0.85	1.76 (1.05-2.95)
	c	0.91 (0.73-1.14)	0.42	1.50 (1.12-2.02)
DHX35-LINC01734				
rs7267944	r	0.91 (0.63-1.32)	0.61	1.17 (0.51-2.68)
	c	1.21 (0.97-1.49)	0.087	2.31 (1.45-3.70)
SPATA13				
rs1220597	r	1.23 (0.82-1.83)	0.32	1.02 (0.63-1.66)
	c	1.19 (0.95-1.50)	0.14	1.38 (1.04-1.83)
FOXA2				
rs1203952	r	1.14 (0.79-1.65)	0.49	1.88 (0.99-3.55)
	c	1.37 (1.11-1.69)	$3.35x10^{-3}$	1.90 (1.24-2.90)
GALNT18				
rs7935113	r	0.48 (0.26-0.87)	0.016	0.53 (0.10-2.75)
	c	1.27 (0.98-1.64)	0.067	1.27 (0.68-2.38)
CYP1A1				
rs1799814	r	1.07 (0.6-1.93)	0.81	2.17 (0.34-13.7)
	c	1.38 (1.01-1.89)	0.046	2.53 (0.64-10.1)
TLR10				
rs11466653	r	0.78 (0.32-1.93)	0.60	-
	c	1.53 (1.01-2.31)	0.043	1.23 (0.08-19.0)
FTO				
rs1121980	r	1.15 (0.76-1.73)	0.52	1.31 (0.81-2.10)
	c	0.98 (0.78-1.23)	0.86	0.73 (0.55-0.96)
rs8050136	r	1.36 (0.74-2.49)	0.32	1.12 (0.53-2.39)

Accepted Manuscript published as ETJ-22-0058.R2. Accepted for publication: 17-Aug-2022

	c	0.97 (0.74-1.26)	0.80	0.70 (0.50-0.97)
RARRES1				
rs7617304	r	0.78 (0.54-1.12)	0.17	1.12 (0.56-2.24)
	c	1.24 (1.00-1.52)	0.045	1.67 (1.13-2.48)
NIS- SLC5A5				
rs35036312	r	0.77 (0.38-1.54)	0.45	1.20 (0.36-4.03)
	c	1.18 (0.93-1.50)	0.18	1.62 (0.99-2.67)
rs4808708	r	1.31 (0.86-1.99)	0.21	1.95 (0.76-5.05)
	c	1.39 (1.10-1.76)	$6.45x10^{-3}$	1.93 (1.14-3.27)
CARD9-SNAPC4				
rs10781500	r	0.97 (0.67-1.40)	0.88	0.90 (0.52-1.56)
	c	0.97 (0.79-1.21)	0.80	0.71 (0.52-0.97)
WDR-SPAG17				
rs4658973	r	1.8 (0.98-3.29)	0.056	1.16 (0.54-2.50)
	c	0.84 (0.65-1.08)	0.18	0.93 (0.66-1.31)
CDKN1B				
rs2066827	r	1.04 (0.59-1.84)	0.88	1.33 (0.47-3.75)
	c	0.96 (0.76-1.23)	0.77	1.78 (0.96-2.96)
MET				
rs1621	r	0.89 (0.49-1.63)	0.72	1.09 (0.46-2.61)
	c	0.84 (0.66-1.08)	0.19	1.43 (0.98-2.09)
IL10				
rs1800896	r	1.16 (0.62-2.17)	0.64	0.75 (0.3-1.86)
	c	1.12 (0.87-1.45)	0.39	0.72 (0.49-1.05)
AXIN2				
rs11655966	r	0.81 (0.44-1.49)	0.50	1.06 (0.33-3.45)
	c	0.97 (0.74-1.27)	0.81	1.76 (0.91-3.39)
PTCSC3-				
rs944289	r	0.63 (0.43-0.93)	0.019	0.41 (0.24-0.69)
	c	0.75 (0.60-0.94)	0.014	0.60 (0.45-0.80)
STAT1				
rs10173099	r	1.05 (0.55-1.98)	0.89	1.24 (0.59-2.61)
	c	1.13 (0.87-1.46)	0.37	1.45 (1.02-2.07)
IL17RB				
rs1025689	r	0.85 (0.58-1.23)	0.38	1.2 (0.72-2.02)
	c	0.86 (0.69-1.08)	0.20	0.84 (0.62-1.13)
IL11RA				
rs1061758	r	1.21 (0.82-1.79)	0.34	2.93 (1.00-8.54)
	c	1.21 (0.97-1.51)	0.095	2.18 (1.17-4.07)
HABP2				
rs7080536	r	1.21 (0.36-4.04)	0.76	-
	c	0.80 (0.45-1.41)	0.43	-
ESR1				
rs2228480	r	0.84 (0.55-1.28)	0.41	0.51 (0.18-1.50)
	c	0.87 (0.68-1.10)	0.24	0.52 (0.25-1.07)
ALOX12				
rs1126667	r	0.89 (0.48-1.63)	0.70	0.65 (0.29-1.49)
	c	1.14 (0.88-1.47)	0.31	0.69 (0.46-1.01)
LINC01465				
rs10877887	r	1.18 (0.80-1.76)	0.40	1.25 (0.75-2.09)

	c	1.25 (1.00-1.56)	0.046	1.34 (1.00-1.81)

 Accepted Manuscript published as ETJ-22-0058.R2. Accepted for publication: 17-Aug-2022

with the dataset 1.

ndex) are provided with their 95% confidence intervals (ORadj; 95% CI). Associations are unc
‹10exp-3).

	Dominant	P_{ass}	Recessive	
P_{ass}	OR_{adj} (95% CI)		OR_{adj} (95% CI)	P_{ass}
$7.76x10^{-4}$	1.53 (1.05-2.22)	0.027	2.30 (1.37-3.84)	$1.51x10^{-3}$
$4.45x10^{-13}$	2.05 (1.66-2.53)	$3.02x10^{-11}$	2.39 (1.78-3.20)	$5.37x10^{-9}$
$3.29x10^{-3}$	1.69 (1.15-2.47)	$7.25x10^{-3}$	1.65 (1.08-2.51)	0.019
$9.98x10^{-11}$	1.89 (1.50-2.37)	$3.72x10^{-8}$	1.94 (1.51-2.49)	$1.79x10^{-7}$
0.067	1.40 (0.99-1.99)	0.060	1.63 (0.88-3.03)	0.12
$4.52x10^{-8}$	1.64 (1.34-2.01)	$1.72x10^{-6}$	2.48 (1.71-3.60)	$1.71x10^{-6}$
0.43	1.01 (0.71-1.43)	0.97	0.75 (0.41-1.38)	0.36
$5.87.0x10^{-4}$	0.75 (0.61-0.91)	$4.33x10^{-3}$	0.59 (0.42-0.84)	$2.83x10^{-3}$
0.64	1.08 (0.75-1.54)	0.68	1.09 (0.68-1.76)	0.71
$4.93x10^{-4}$	1.34 (1.09-1.64)	$6.27x10^{-3}$	1.53 (1.16-2.02)	$2.76x10^{-3}$
0.11	0.98 (0.68-1.41)	0.90	1.86 (0.96-3.62)	0.066
0.087	0.75 (0.61-0.92)	$6.16x10^{-3}$	0.77 (0.51-1.17)	0.22
0.59	0.73 (0.49-1.09)	0.12	1.62 (0.37-7.14)	0.52
0.36	0.70 (0.56-0.88)	$2.06x10^{-3}$	0.78 (0.38-1.61)	0.51
0.020	1.18 (0.81-1.70)	0.39	1.8 (1.18-2.75)	$6.6x10^{-3}$
$2.83x10^{-4}$	1.24 (1.01-1.54)	0.044	1.61 (1.25-2.08)	$2.01x10^{-4}$
0.033	1.19 (0.81-1.74)	0.37	1.72 (1.09-2.69)	0.019
$6.55x10^{-3}$	1.04 (0.84-1.29)	0.70	1.59 (1.23-2.06)	$4.70x10^{-4}$
0.71	0.94 (0.66-1.34)	0.72	1.21 (0.54-2.74)	0.65
$4.64x10^{-4}$	1.31 (1.07-1.61)	$8.65x10^{-3}$	2.16 (1.36-3.44)	$1.09x10^{-3}$
0.93	1.16 (0.79-1.69)	0.45	0.9 (0.6-1.37)	0.63
0.025	1.25 (1.00-1.54)	0.046	1.24 (0.97-1.59)	0.081
0.052	1.25 (0.89-1.77)	0.20	1.79 (0.96-3.34)	0.066
$3.12x10^{-3}$	1.44 (1.17-1.76)	$4.16x10^{-4}$	1.68 (1.11-2.54)	0.015
0.45	0.48 (0.27-0.86)	0.014	0.64 (0.13-3.27)	0.59
0.45	1.27 (0.99-1.62)	0.056	1.18 (0.64-2.20)	0.59
0.41	1.14 (0.65-2.00)	0.64	2.16 (0.34-13.62)	0.41
0.19	1.42 (1.04-1.93)	0.026	2.44 (0.61-9.69)	0.21
-	0.68 (0.28-1.62)	0.38	-	-
0.88	1.52 (1.01-2.29)	0.043	1.18 (0.08-inf)	0.91
0.27	1.20 (0.82-1.75)	0.35	1.20 (0.80-1.80)	0.37
0.027	0.89 (0.72-1.11)	0.30	0.74 (0.58-0.94)	0.014
0.76	1.28 (0.73-2.24)	0.38	0.96 (0.48-1.91)	0.90

0.034	0.88 (0.69-1.13)	0.31	0.71 (0.53-0.96)	0.024
0.75	0.82 (0.58-1.16)	0.26	1.23 (0.62-2.44)	0.55
0.011	1.30 (1.06-1.58)	0.011	1.53 (1.04-2.24)	0.032
0.76	0.83 (0.43-1.59)	0.57	1.33 (0.41-4.33)	0.64
0.056	1.23 (0.98-1.55)	0.077	1.52 (0.94-2.48)	0.091
0.17	1.37 (0.92-2.06)	0.12	1.78 (0.70-4.56)	0.23
0.015	1.45 (1.15-1.82)	$1.36x10^{-3}$	1.72 (1.02-2.89)	0.042
0.71	0.95 (0.68-1.35)	0.79	0.91 (0.54-1.54)	0.73
0.029	0.90 (0.74-1.10)	0.32	0.72 (0.53-0.96)	0.026
0.71	1.59 (0.91-2.78)	0.10	0.83 (0.41-1.67)	0.61
0.66	0.86 (0.67-1.10)	0.22	1.03 (0.76-1.40)	0.86
0.59	1.09 (0.63-1.86)	0.76	1.31 (0.48-3.56)	0.60
0.054	1.07 (0.85-1.35)	0.56	1.75 (0.98-2.98)	0.051
0.83	0.94 (0.53-1.65)	0.83	1.16 (0.51-2.61)	0.72
0.061	0.95 (0.75-1.21)	0.68	1.53 (0.91-2.23)	0.054
0.54	1.06 (0.59-1.91)	0.85	0.69 (0.30-1.59)	0.39
0.089	1.02 (0.80-1.30)	0.88	0.68 (0.47-1.01)	0.051
0.92	0.85 (0.48-1.5)	0.58	1.13 (0.35-3.62)	0.84
0.091	1.04 (0.80-1.35)	0.76	1.78 (0.93-3.40)	0.082
$8.00x10^{-4}$	0.56 (0.39-0.80)	$1.7x10^{-3}$	0.53 (0.33-0.85)	$9.00x10^{-3}$
$5.33x10^{-4}$	0.71 (0.57-0.87)	$1.11x10^{-3}$	0.71 (0.55-0.91)	$8.04x10^{-3}$
0.58	1.11 (0.62-1.99)	0.72	1.21 (0.62-2.33)	0.58
0.041	1.20 (0.94-1.53)	0.14	1.36 (0.98-1.88)	0.066
0.49	0.92 (0.64-1.31)	0.64	1.32 (0.82-2.12)	0.25
0.24	0.86 (0.70-1.06)	0.15	0.91 (0.69-1.19)	0.49
0.049	1.31 (0.90-1.90)	0.16	2.77 (0.96-8.03)	0.060
0.015	1.28 (1.03-1.59)	0.027	2.06 (1.11-3.84)	0.022
-	1.21 (0.36-4.04)	0.76	-	-
-	0.80 (0.45-1.41)	0.43	-	-
0.22	0.79 (0.53-1.18)	0.25	0.53 (0.18-1.55)	0.25
0.076	0.83 (0.66-1.04)	0.11	0.54 (0.26-1.11)	0.091
0.31	0.82 (0.46-1.46)	0.51	0.70 (0.34-1.47)	0.35
0.051	1.01 (0.79-1.28)	0.96	0.67 (0.42-1.02)	0.052
0.39	1.20 (0.83-1.74)	0.33	1.14 (0.72-1.81)	0.58

Accepted Manuscript published as ETJ-22-0058.R2. Accepted for publication: 17-Aug-2022

0.052	1.28 (1.04-1.57)	0.022	1.19 (0.90-1.56)	0.22

derlined when statistically

Chi-sq	P-trend
28.0	1.20×10^{-7}
68.4	$<10^{-15}$
15.2	9.71×10^{-5}
18.1	2.07×10^{-5}
12.05	5.19×10^{-4}
38.5	5.37×10^{-10}
2.17	0.14
15.7	7.35×10^{-5}
0.69	0.41
11.8	5.88×10^{-4}
0.29	0.59
6.77	9.27×10^{-3}
0.076	0.78
9.15	2.49×10^{-3}
4.09	0.043
17.7	2.62×10^{-5}
3.98	0.046
10.1	1.48×10^{-3}
0.27	0.60
10.7	1.09×10^{-3}
0.003	0.95
7.00	8.14×10^{-3}
6.42	0.011
20.3	6.49×10^{-6}
4.12	0.042
4.95	0.0261
1.00	0.32
10.3	1.36×10^{-3}
0.60	0.44
6.51	0.0107
1.67	0.20
1.95	0.162
0.72	0.40

Accepted Manuscript published as ETJ-22-0058.R2. Accepted for publication: 17-Aug-2022

4.52	0.0336
1.75	0.19
10.6	1.14×10^{-3}
0.62	0.43
3.17	0.0749
1.82	0.18
7.33	6.78×10^{-3}
0.15	0.70
3.16	0.0757
0.30	0.58
2.08	0.150
0.63	0.42
1.78	0.182
0.014	0.91
0.446	0.504
0.22	0.64
0.816	0.366
0.57	0.45
1.30	0.254
18.06	2.14×10^{-5}
21.9	2.81×10^{-6}
1.03	0.31
3.73	0.066
0.10	0.75
3.94	0.0473
2.32	0.13
6.82	9.02×10^{-3}
-	-
-	-
0.063	0.80
1.36	0.243
0.36	0.24
0.367	0.544
1.42	0.23

5.31	0.0212

Accepted Manuscript published as ETJ-22-0058.R2. Accepted for publication: 17-Aug-2022

Supplementary table S4. Classification metrics of Gaussian Naïve Bayes classifier on all datase

Metric	Training Set	Test Set	Validation Set
NPV	0.51	0.46	0.43
PPV	0.74	0.76	0.81
Sensitivity	0.46	0.46	0.43
Specificity	0.77	0.76	0.81
Accuracy	59%	57%	56%
F1-score	0.57	0.57	0.56
F0.5-score	0.66	0.67	0.69
F2-score	0.50	0.50	0.48

ets.

 Accepted Manuscript published as ETJ-22-0058.R2. Accepted for publication: 17-Aug-2022

Supplementary Table S5: method of calculation of unweighted and weighted polygenic risk score (PRS). The unweighted PRS was built by summing the total number of risk alleles for each subject (attributing the value of 1 to each risk allele). In the weighted PRS, all the SNPs contribute to the total PRS according to their association with the risk. It was built by assigning to each genotype the relative OR obtained in the GWAS. Then the ORs were multiplied.

	Genetic locus	1	2	3	4	5	6	7	10	11	12	13	14	15	16	17	18	19	20	21	
	Risk allele	T	C	T	T	T	A	A	T	T	C	C	A	G	T	A	A	G	G	A	
Subject																					
1		CC	CC	TT	CC	AG	CC	AA	AT	CC	CC	CT	GA	GA	GT	AA	AG	GG	AA	AA	
2		TT	TC	TT	CC	AA	CC	GA	AT	CC	CC	CC	GG	GG	GT	AG	AG	GA	AA	AT	
3		CC	CC	TC	CC	AA	CC	GG	AA	CC	CA	CC	GA	GG	TT	AA	GG	GG	AA	AA	
4		CT	TC	TT	CC	AA	CC	GA	AA	CT	CC	CC	GG	GA	GT	AG	GG	GA	AA	AA	
5		CT	TT	TT	CC	AA	CC	GA	AA	CC	CA	TT	GA	GG	GT	AA	AA	GG	AA	AT	
6		TT	TC	TC	CC	AA	CC	GG	AT	CC	CA	CT	GG	GG	GT	AA	GG	GG	AA	AA	
7		CT	TC	TC	CT	AA	CC	AA	AA	CC	CA	CC	GG	GG	GT	AG	AG	AA	AA	AA	
8		CT	TC	TT	CC	AG	CC	GA	AA	CC	CA	CC	GG	GA	GG	GG	AA	GG	AG	AA	
9		CT	TC	TC	CC	AA	CC	GA	AT	CC	CC	CC	GA	GA	GT	AA	AA	GG	AA	AA	
Unweighted PRS																					**PRS (is the summ)**
Subject																					
1		0	2	2	0	0	0	2	1	0	2	1	1	1	1	2	1	2	0	2	20
2		2	1	2	0	0	0	1	1	0	2	2	0	2	1	1	1	1	0	1	18
3		0	2	1	0	0	0	0	0	0	1	2	1	2	2	2	0	2	0	2	17
4		1	1	2	0	0	0	1	0	1	2	2	0	1	1	1	0	1	0	2	16
5		1	0	2	0	0	0	1	0	0	1	0	1	2	1	2	2	2	0	1	16
6		2	1	1	0	0	0	0	1	0	1	1	0	2	1	2	0	2	0	2	16
7		1	1	1	1	0	0	2	0	0	1	2	0	2	1	1	1	0	0	2	16
8		1	1	2	0	0	0	1	0	0	1	2	0	1	0	0	2	2	1	2	16
9		1	1	1	0	0	0	1	1	0	2	2	1	1	1	2	2	2	0	2	20
Weighted PRS																					
OR [a]		1.20	1.18	1.20	1.12	1.17	1.26	0.95	1.12	1.25	1.03	1.13	1.02	1.13	0.97	1.29	1.05	1.07	1.34	0.91	
OR [b]		1.31	1.43	1.24	1.28	1.41	1.67	1.17	1.40	1.60	1.37	1.56	1.21	1.24	1.21	1.31	1.27	1.31	1.95	1.91	
Subject																					**PRS (is the product)**
1		1	1.43	1.24	1	1	1	1.17	1.12	1	1.37	1.13	1.02	1.13	0.97	1.31	1.05	1.31	1	1.91	13.84
2		1.31	1.18	1.24	1	1	1	0.95	1.12	1	1.37	1.56	1	1.24	0.97	1.29	1.05	1.07	1	0.91	6.91
3		1	1.43	1.2	1	1	1	1	1	1	1.03	1.56	1.02	1.24	1.21	1.31	1	1.31	1	1.91	13.83
4		1.2	1.18	1.24	1	1	1	0.95	1	1.25	1.37	1.56	1	1.13	0.97	1.29	1	1.07	1	1.91	12.88
5		1.2	1	1.24	1	1	1	0.95	1	1	1.03	1	1.02	1.24	0.97	1.31	1.27	1.31	1	0.91	3.54
6		1.31	1.18	1.2	1	1	1	1	1.12	1	1.03	1.13	1	1.24	0.97	1.31	1	1.31	1	1.91	9.53
7		1.2	1.18	1.2	1.12	1	1	1.17	1	1	1.03	1.56	1	1.24	0.97	1.29	1.05	1	1	1.91	11.13
8		1.2	1.18	1.24	1	1	1	0.95	1	1	1.03	1.56	1	1.13	1	1	1.27	1.31	1.34	1.91	12.90
9		1.2	1.18	1.2	1	1	1	0.95	1.12	1	1.37	1.56	1.02	1.13	0.97	1.31	1.27	1.31	1	1.91	17.98

a= heterozygotes vs common homozygotes

b= rare homozygotes vs common homozygotes