



Article

# Genome-Wide Association Analysis and Genomic Prediction of Thyroglobulin Plasma Levels

Nikolina Pleić <sup>1</sup>, Mirjana Babić Leko <sup>1</sup>, Ivana Gunjača <sup>1</sup>, Thibaud Boutin <sup>2</sup>, Vesela Torlak <sup>3</sup>, Antonela Matana <sup>1,†</sup>, Ante Punda <sup>3</sup>, Ozren Polašek <sup>4</sup>, Caroline Hayward <sup>2</sup> and Tatijana Zemunik <sup>1,\*</sup>

<sup>1</sup> Department of Medical Biology, School of Medicine, University of Split, Šoltanska 2, 21000 Split, Croatia; npleic@mefst.hr (N.P.); mbabic@mefst.hr (M.B.L.); igunjaca@mefst.hr (I.G.); antonela.matana@gmail.com (A.M.);

<sup>2</sup> MRC Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh EH4 2XU, UK; tboutin@ed.ac.uk (T.B.); caroline.hayward@ed.ac.uk (C.H.)

<sup>3</sup> Department of Nuclear Medicine, University Hospital Split, Spinčičeva 1, 21000 Split, Croatia; veselakbsplit@yahoo.com (V.T.); ante.punda@mefst.hr (A.P.)

<sup>4</sup> Department of Public Health, School of Medicine, University of Split, Šoltanska 2, 21000 Split, Croatia; opolasek@gmail.com

\* Correspondence: tzemunik@mefst.hr; Tel.: +385-2155-7888

† Current address: The University Department of Health Studies, University of Split, 21000 Split, Croatia.

**Abstract:** Thyroglobulin (Tg) is an iodoglycoprotein produced by thyroid follicular cells which acts as an essential substrate for thyroid hormone synthesis. To date, only one genome-wide association study (GWAS) of plasma Tg levels has been performed by our research group. Utilizing recent advancements in computation and modeling, we apply a Bayesian approach to the probabilistic inference of the genetic architecture of Tg. We fitted a Bayesian sparse linear mixed model (BSLMM) and a frequentist linear mixed model (LMM) of 7,289,083 variants in 1096 healthy European-ancestry participants of the Croatian Biobank. Meta-analysis with two independent cohorts (total  $n = 2109$ ) identified 83 genome-wide significant single nucleotide polymorphisms (SNPs) within the *ST6GAL1* gene ( $p < 5 \times 10^{-8}$ ). BSLMM revealed additional association signals on chromosomes 1, 8, 10, and 14. For *ST6GAL1* and the newly uncovered genes, we provide physiological and pathophysiological explanations of how their expression could be associated with variations in plasma Tg levels. We found that the SNP-heritability of Tg is 17% and that 52% of this variation is due to a small number of 16 variants that have a major effect on Tg levels. Our results suggest that the genetic architecture of plasma Tg is not polygenic, but influenced by a few genes with major effects.

**Keywords:** genome-wide association study; thyroglobulin; thyroid; ST6GAL1; LMM; BSLMM



**Citation:** Pleić, N.; Babić Leko, M.; Gunjača, I.; Boutin, T.; Torlak, V.; Matana, A.; Punda, A.; Polašek, O.; Hayward, C.; Zemunik, T. Genome-Wide Association Analysis and Genomic Prediction of Thyroglobulin Plasma Levels. *Int. J. Mol. Sci.* **2022**, *23*, 2173. <https://doi.org/10.3390/ijms23042173>

Academic Editor: Concetta Ambrosino

Received: 25 January 2022

Accepted: 12 February 2022

Published: 16 February 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Thyroglobulin (Tg) is the most abundant protein produced by the thyroid gland. This 660 kDa iodoglycoprotein serves as a storehouse of thyroid hormones since Tg proteolysis releases thyroxine (T4) and triiodothyronine (T3) [1]. Tg is synthesized in thyrocytes. Following the post-translational modifications occurring in the rough endoplasmic reticulum and the Golgi apparatus, Tg is released into the follicular lumen where Tg iodination and hormone production occur [2,3]. Mature Tg is then transferred back to the thyrocytes by endocytosis. In the thyrocytes, Tg proteolysis occurs and thyroid hormones are released into the bloodstream on the basolateral membrane [1]. Some portion of intact Tg (mostly poorly sialylated or iodinated) can be transferred by transcytosis from the follicular lumen to the bloodstream [4]. In addition to transcytosis, Tg can be released into the blood from disrupted follicles. Moreover, plasma Tg levels are increased in thyroid pathology, and plasma Tg levels have been shown to correlate with thyroid mass [5]. A twin study showed that the observed variability in serum Tg levels has a strong genetic component [6].

Approximately 10% of Tg molecular mass is glycosylated [7]. The glycosylation of Tg is crucial for the synthesis of thyroid hormones because it has been shown that unglycosylated Tg has no potential for the synthesis of thyroid hormones [8]. Glycosylation is also important for Tg folding, iodination, trafficking and immunoreactivity [1]. Sialylation is a late post-translational modification of Tg that occurs in the Golgi apparatus [1]. ST6GAL1 ( $\beta$ -galactoside  $\alpha$ -2,6-sialyltransferase), also known as sialyltransferase 1, catalyzes the addition of  $\alpha$ -2,6 bound sialic acid to N-glycosylated proteins [9]. It is involved in the sialylation of Tg since  $\alpha$ -2,6 bound sialic acid residues are detected at Tg [10,11]. Both ST6GAL1 mRNA [12–14] and protein [13,14] were detected in the thyroid gland. This membrane-bound enzyme is mainly found in the Golgi apparatus [15]. Sialylation is important for many Tg functions: immunoreactivity, autoregulation, and recycling. The desialylation of Tg increases its immunoreactivity [16,17]. In addition, poorly iodinated or sialylated Tg has a higher potential to trigger Tg-mediated signaling [18]. Sialylation also affects Tg recycling because it is important for binding Tg to its transmembrane receptor [17]. The importance of Tg sialylation for its proper functioning is evident from the case of a patient with congenital goiter with hypothyroidism. This patient had severely hyposialylated Tg and insufficient  $\alpha$ -2,6 sialyltransferase activity [10].

Our recent genome-wide association study (GWAS) showed an association of 16 variants within the *ST6GAL1* gene with plasma Tg levels in healthy individuals [19]. This was the first GWAS to investigate genes associated with plasma Tg levels. The linear mixed model (LMM) used in our study has become a standard for genome-wide association mapping because it efficiently controls for both population structure and relatedness among individuals. However, LMMs as well as other frequentist methods only test one single nucleotide polymorphism (SNP) at a time. On the other hand, methods that relate phenotypic variation to multiple genetic variants simultaneously could further increase the power to detect causal variants. Multiple SNP modeling extensions of the standard LMM have been proposed from a Bayesian perspective by considering alternative prior distributions on the genetic effects. In the current study, we included an additional 1096 individuals and conducted association mapping using both frequentist and Bayesian approaches, as well as SNP heritability estimation and genomic prediction using Bayesian approaches. We aimed to replicate the significant findings to further confirm the association of the *ST6GAL1* gene with plasma Tg levels in healthy individuals. Additionally, we sought to elucidate the genetic architecture of Tg by using Bayesian multi-SNP approaches. Finally, we meta-analyzed our new GWA results with our previously published GWA results in a combined dataset of 2190 individuals. The outcome of such a comprehensive approach will be the generation of new knowledge on the genetic background of Tg that will lead to a better understanding of the biological pathways related to thyroid function.

## 2. Results

### 2.1. Genome-Wide Association Analyses

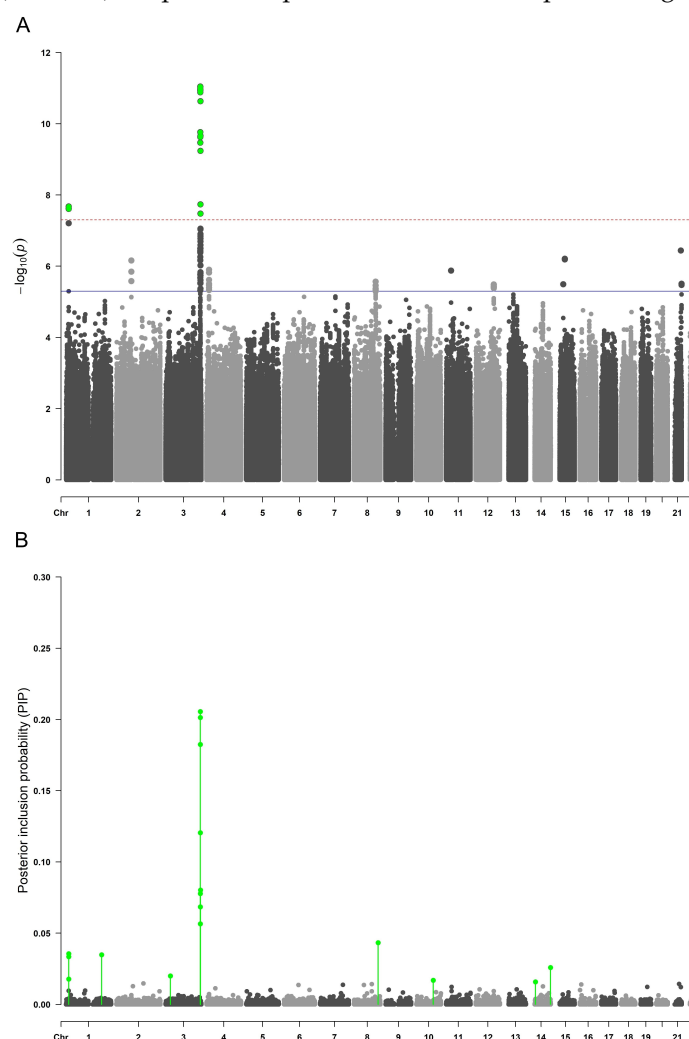
In the new LMM association analysis, a total of 18 SNPs reached genome-wide significance. Of the significantly associated SNPs, 15 were located within the *ST6GAL1* gene on chromosome 3, and 3 were located within the *PDPN* gene on chromosome 1 (Table 1). Among the 15 SNPs within the *ST6GAL1* gene that reached genome-wide significance, 11 were replications of our previously published discovery phase genome-wide significant results. The other five genome-wide significant variants from the discovery phase were also replicated at the  $1 \times 10^{-4}$   $p$ -value threshold (Supplementary Table S4).

**Table 1.** SNPs passing genome-wide significance threshold ( $5 \times 10^{-8}$ ) in the single-SNP LMM analysis and their corresponding PIPs from the multi-SNP BSLMM analysis of cohorts Korcula2 and Korcula3.

SNP	Chr	Position	Gene	Ref. Allele	Effect Allele	EAF	Single-SNP LMM Analysis in Cohorts Korcula2 and Korcula3 $\beta$ ( $p$ -Value)	Multi-SNP BSLMM Analysis in Cohorts Korcula2 and Korcula3 $\beta$ (PIP)
rs10937280	3	186738033	<i>ST6GAL1</i>	G	A	0.35	−0.31 ( $9.09 \times 10^{-12}$ )	−0.29 (0.21)
rs5001409	3	186735690	<i>ST6GAL1</i>	A	C	0.35	−0.31 ( $9.44 \times 10^{-12}$ )	−0.295 (0.07)
rs9863411	3	186737820	<i>ST6GAL1</i>	C	T	0.35	−0.31 ( $1.06 \times 10^{-11}$ )	−0.283 (0.2)
rs7634389	3	186738421	<i>ST6GAL1</i>	T	C	0.35	−0.31 ( $1.12 \times 10^{-11}$ )	−0.292 (0.08)
rs967367	3	186734466	<i>ST6GAL1</i>	G	A	0.35	−0.31 ( $1.15 \times 10^{-11}$ )	−0.29 (0.12)
rs3821819	3	186732725	<i>ST6GAL1</i>	G	A	0.35	−0.31 ( $1.31 \times 10^{-11}$ )	−0.292 (0.06)
rs4686838	3	186743053	<i>ST6GAL1</i>	A	G	0.45	−0.3 ( $2.33 \times 10^{-11}$ )	−0.27 (0.08)
rs10212190	3	186731157	<i>ST6GAL1</i>	A	T	0.34	−0.29 ( $1.73 \times 10^{-10}$ )	−0.28 (0.003)
rs4012172	3	186741511	<i>ST6GAL1</i>	C	T	0.36	−0.29 ( $2.19 \times 10^{-10}$ )	−0.27 (0.0003)
rs3872724	3	186741221	<i>ST6GAL1</i>	C	T	0.37	−0.28 ( $2.37 \times 10^{-10}$ )	−0.27 (0.001)
rs3872723	3	186741131	<i>ST6GAL1</i>	C	T	0.36	−0.28 ( $3.4 \times 10^{-10}$ )	0 (0)
rs28674898	3	186744563	<i>ST6GAL1</i>	G	A	0.39	0.28 ( $5.81 \times 10^{-10}$ )	−0.28 (0.003)
rs4686844	3	186765135	<i>ST6GAL1</i>	G	A	0.56	−0.25 ( $1.83 \times 10^{-10}$ )	−0.15 (0.0007)
rs78946539	1	13921500	<i>PDPN</i>	A	G	0.04	−0.63 ( $2.1 \times 10^{-8}$ )	−0.51 (0.03)
rs143154928	1	13921447	<i>PDPN</i>	G	A	0.04	−0.63 ( $2.32 \times 10^{-8}$ )	−0.5 (0.03)
rs12566684	1	13922117	<i>PDPN</i>	A	G	0.04	−0.64 ( $2.46 \times 10^{-8}$ )	−0.5 (0.02)
rs257104	3	186775807	<i>ST6GAL1</i>	G	A	0.4	0.24 ( $3.33 \times 10^{-8}$ )	0.17 (0.002)

Statistical analyses were performed with GEMMA LMM and BSLMM.  $p$ -values  $< 5 \times 10^{-8}$  are genome-wide significant. SNPs are sorted by ascending LMM  $p$ -value. BSLMM, Bayesian sparse linear mixed model; Chr, chromosome; EAF, effect allele frequency; LMM, linear mixed model; PIP, posterior inclusion probability; SNP, single nucleotide polymorphism.

In the BSLMM association analysis, 16 SNPs were identified as having a major sparse effect on plasma Tg levels and these variants were estimated to have a sparse effect in  $\geq 1.6\%$  of BSLMM chain iterations (i.e., posterior inclusion probability (PIP)  $\geq 0.016$ ) (Supplementary Table S3). Moreover, the top four SNPs were identified as having a sparse effect on Tg levels in more than 10% of chain iterations (PIP  $> 0.1$ ) and all were located within the *ST6GAL1* gene. There was a complete overlap in the significant results identified in single-SNP LMM association analysis and multi-SNP BSLMM analysis for the variants located within the *ST6GAL1* gene on chromosome 3 and *PDPN* gene on chromosome 1 (Table 1). The BSLMM approach uncovered additional association signals on chromosome 8 (rs10283166—*PVT1* gene intron variant), chromosome 14 (rs35862113—*MARK3* gene intron variant, rs61972442—*OR6J1* gene intron variant), chromosome 3 (rs1631354—*RARB* gene intron variant), and chromosome 10 (rs11202702—*RNLS* gene intron variant). Results from the single-SNP association analysis (LMM) and the multi-SNP association analysis (BSLMM) are plotted in parallel in Manhattan plots in Figure 1.



**Figure 1.** Manhattan plots of single-SNP and multi-SNP association mapping in cohorts Korcula2 and Korcula3. **(A)** Manhattan plot of single-SNP LMM analysis. The x axis represents the chromosomal position of SNPs and the y axis represents their  $-\log_{10}(p\text{-values})$  obtained by the LMM analysis. Each dot on the Manhattan plot signifies an SNP. Because the strongest associations have the smallest  $p$ -values (e.g.,  $10^{-12}$ ), their negative logarithms will be the greatest (e.g., 12). The red horizontal line indicates the genome-wide significance threshold ( $p = 5 \times 10^{-8}$ ), while the blue horizontal line indicates the suggestive threshold of significance ( $p = 5 \times 10^{-6}$ ). **(B)** Manhattan plot of multi-SNP BSLMM analysis. The x axis represents the chromosomal position of SNPs, and the y axis represents their posterior inclusion probabilities (PIPs) obtained by the BSLMM analysis.

## 2.2. SNP Heritability Estimation

In our previous work [19], the top rs4012172 SNP was estimated to explain 3.19% of the variance in Tg levels. In the current study, we estimated the proportion of variance in phenotypes explained by all available genotypes (PVE) or the “chip heritability”, as well as the proportion of genetic variance explained by variants with major effect (PGE). The PVE estimate from the BSLMM with 7,289,083 SNPs indicated that 17% of the variation in plasma Tg levels was explained by all available genotypes and that 52% (PGE) of this variation was due to 16 SNPs with relatively large phenotypic effects. These results describe the genetic architecture of plasma Tg and imply that it is not purely polygenic but rather favors the sparse assumption on the variant effects. Means, medians and 95% equal tail posterior probability intervals (95% ETPPIs) of the hyperparameters estimated from the BSLMM are reported in Supplementary Table S5.

## 2.3. Genetic Prediction of Thyroglobulin Levels (Polygenic Score PGS Analysis)

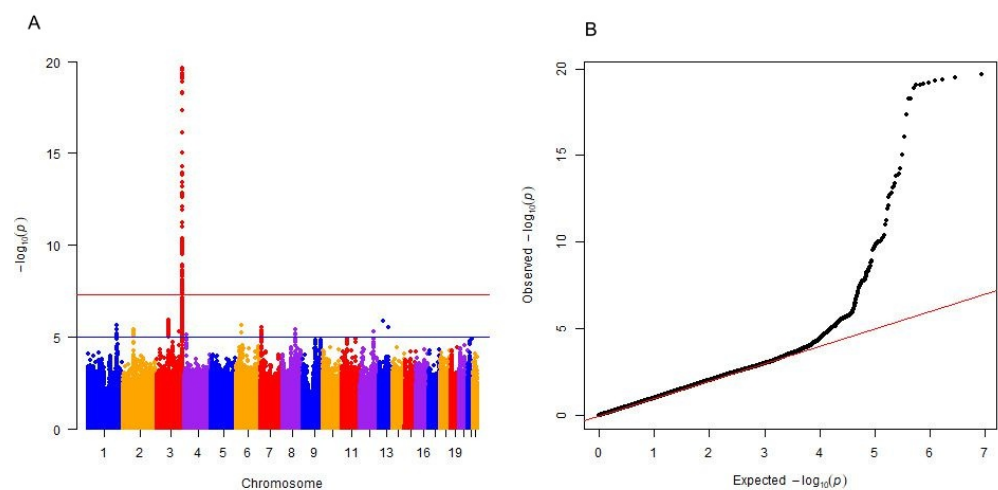
To measure the prediction performance, we calculated the correlation coefficient of predicted and observed values in the test data. Keeping in mind that the PVE was estimated to be 0.17 by the BSLMM, 0.17 was considered as a theoretical upper bound for the accuracy of the predictive model. The Pearson’s correlation coefficient was equal to  $-0.05$  (95% CI  $[-0.1, 0.0004]$ ) with a  $p$ -value of 0.052. This result implies that we have constructed a genomic predictor of plasma Tg levels which, with the inclusion of additional training and test data, is expected to pass the 5% statistical significance threshold.

## 2.4. Meta-Analysis

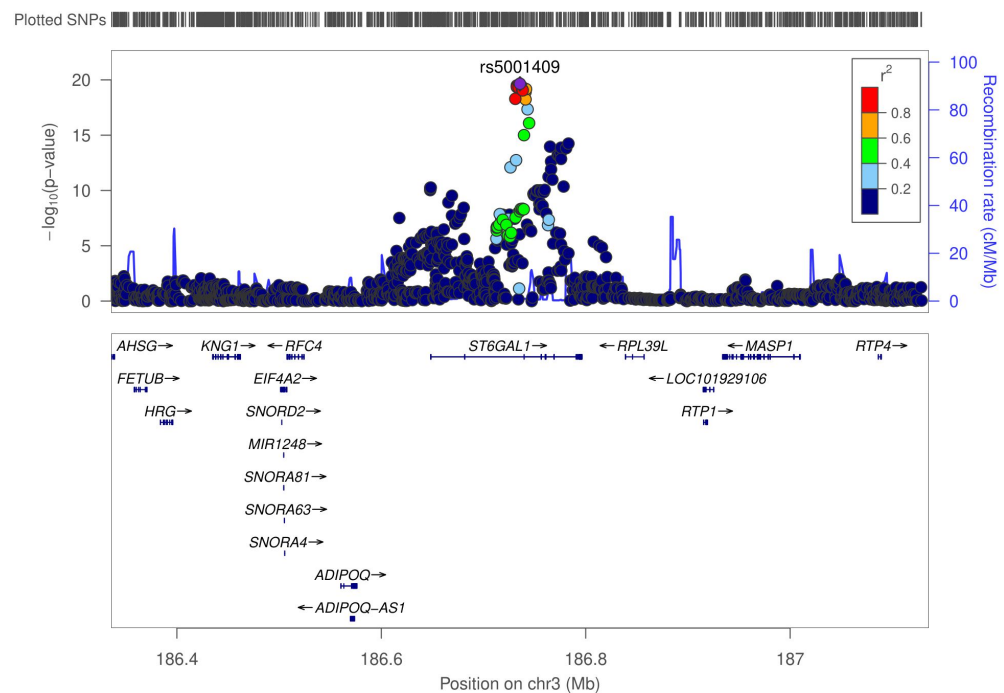
To attain the largest available sample size for this study, the discovery and replication datasets were meta-analyzed in order to uncover additional signals hidden in the separated discovery and replication analyses due to a lack of power. There was little evidence for population stratification at the replication-level ( $\lambda_{Korcula2\&3} = 1.004$ ) or meta-analysis level ( $\lambda = 1.029$ ). In the meta-analysis phase, 83 SNPs within the *ST6GAL1* gene on chromosome 3 reached genome-wide significance (Figure 2 and Supplementary Table S6). The most significant SNP was rs5001409 ( $p = 1.85 \times 10^{-20}$ ). The regional association plot of the *ST6GAL1* region is shown in Figure 3. The minor C allele (MAF = 0.38) of the rs5001409 was associated with lower Tg levels ( $\beta = -0.297$ , SE = 0.03). Effect sizes were in the same direction in all datasets. The forest plot of the effect sizes is shown in Supplementary Figure S1.

## 2.5. Colocalization Analysis

Our analysis supports a strong colocalization of GWAS signals with eQTLs of the *ST6GAL1* gene in thyroid tissue with an SS  $p$ -value of  $1 \times 10^{-7}$ . The colocalization analysis is visualized in Figure 4. According to the GTEx portal, the most significantly associated SNP, rs5001409, was also strongly associated with the expression of the *ST6GAL1* gene in the thyroid tissue ( $p = 1.7 \times 10^{-18}$ ). The association is visualized in the violin plot (Supplementary Figure S2). A normalized effect size (NES) is defined as the slope of the linear regression and is computed as the effect of the alternative allele (C allele) relative to the reference allele (A allele) in the human reference genome (i.e., the eQTL effect allele is the ALT allele). The NES of the C allele at rs5001409 was  $-0.33$ , while the median normalized expression of the *ST6GAL1* gene was 0.1952 for genotype AA,  $-0.0174$  for genotype AC and  $-0.5219$  for genotype CC.

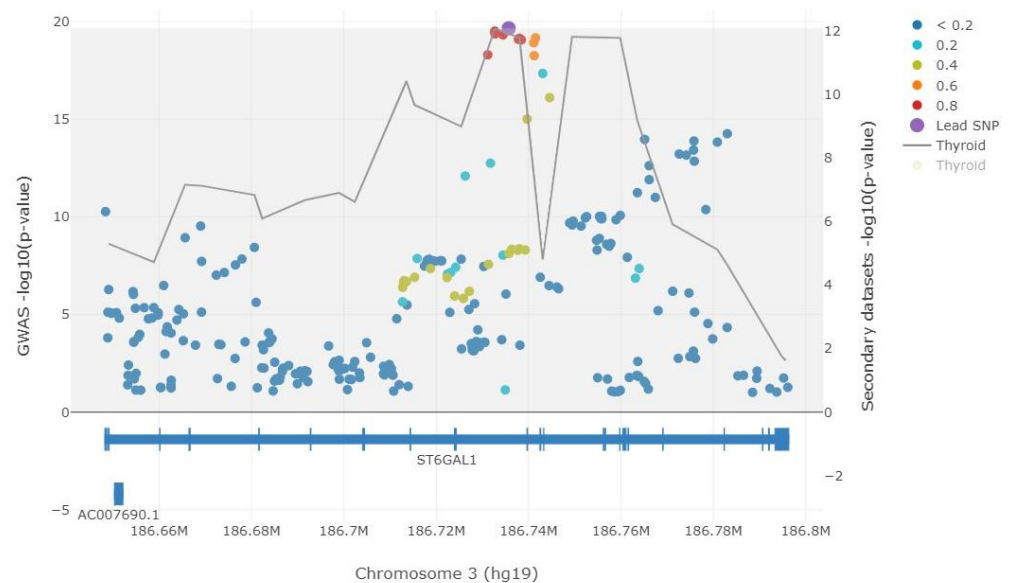


**Figure 2.** Manhattan plot and quantile–quantile (Q-Q) plot of the meta-analysis results for thyroglobulin (Tg) levels. **(A)** Manhattan plot of single nucleotide polymorphisms (SNP) for Tg levels. The x axis represents the chromosomal position of SNPs and the y axis represents their  $-\log_{10}(p\text{-values})$  obtained by combined analysis. Each dot on the Manhattan plot signifies an SNP. The red horizontal line indicates the genome-wide significance threshold ( $p = 5 \times 10^{-8}$ ), while the blue horizontal line indicates the suggestive threshold of significance ( $p = 5 \times 10^{-6}$ ). **(B)** In the Q-Q plot, we see a strong deviation from the null distribution (the distribution of  $p$ -values under the null hypothesis of no true association is indicated by the red line).



**Figure 3.** Regional association plot of the *ST6GAL1* region. The most significant SNP (rs5001409) is shown in purple. The colors of the circles denote their correlations (LD  $r^2$ ) with the top SNP (lead SNP) in purple, high LD SNPs with  $r^2 \geq 0.8$  in red, orange for  $0.8 > r^2 \geq 0.6$ , green for  $0.6 > r^2 \geq 0.4$ , light blue for  $0.4 > r^2 \geq 0.2$  and dark blue for  $r^2 < 0.2$ . The figure was generated using the LocusZoom tool [20].





**Figure 4.** Colocalization analysis of thyroglobulin GWAS signals with eQTL signals of *ST6GAL1* gene in thyroid tissue. Filled circles represent thyroglobulin GWAS  $-\log_{10}(p\text{-values})$  (left  $y$  axis). The rs5001409 SNP was defined as the lead SNP and is presented in purple. The LD information is similar to LocusZoom. The LD information was computed from the European 1000 Genomes subset (phase 1, version 3) [21] in reference to the lead SNP. The gray line represents the eQTL signals and traces the lowest  $p$ -value (right  $y$  axis, showing  $-\log_{10}(p\text{-values})$ ). Gene track information is from GENCODE v19 (hg19 coordinates). The figure was generated using the LocusFocus tool [22].

### 3. Discussion

This study confirmed the results of our recent discovery GWAS on the association of the *ST6GAL1* gene with Tg plasma levels in healthy individuals [19]. In the meta-analysis, we confirmed 16 variants within the *ST6GAL1* gene previously associated with plasma Tg levels [19] and detected an additional 67 variants within the *ST6GAL1* gene that were associated with plasma Tg levels. The strongest association with plasma Tg levels was observed for the *ST6GAL1* gene rs5001409 SNP ( $p = 1.85 \times 10^{-20}$ ). The C allele of this polymorphism was associated with lower plasma Tg levels. The highest expression of the *ST6GAL1* gene was found in the liver, lymph node, spleen, thyroid, and prostate tissue [23]. According to the GTEx portal, the strongest eQTL signals for the lower expression of the *ST6GAL1* gene in thyroid tissue are rs967367, rs3821819, rs10937280, rs17776120 and our top SNP, rs5001409, with an expression  $p$ -value of  $1.7 \times 10^{-18}$ . The top six eQTL signals were also in the top seven signals associated with lower Tg levels in our meta-analysis. Additionally, these SNPs are in high LD with our top rs5001409 variant. This overlap was further confirmed by our colocalization analysis. We offer several explanations of how a decreased *ST6GAL1* expression may be associated with decreased plasma Tg levels. The first possibility is the association of *ST6GAL1* and Tg via the Wnt/ $\beta$ -catenin signaling pathway. *ST6GAL1* activates the Wnt/ $\beta$ -catenin signaling pathway through the PI3K/Akt/GSK-3 $\beta$  signaling pathway [24]. Lower *ST6GAL1* expression leads to a lower activation of the PI3K/Akt/GSK-3 $\beta$  signaling pathway, resulting in the lower activation of the Wnt/ $\beta$ -catenin signaling pathway. Because the Wnt/ $\beta$ -catenin signaling pathway activates the expression of thyroid transcription factor 1 (TTF-1)[25] (a transcription factor involved in TG transcription [26]), the lower activation of this pathway leads to lower levels of TTF-1 and consequently lower Tg levels. The second possibility of *ST6GAL1* and Tg association is via the thyroid-stimulating hormone (TSH) receptor. Namely, *ST6GAL1* adds sialic acid to the TSH receptor [27]. The sialylation of the TSH receptor increases the level of intracellular cAMP [28] (increased concentration of intracellular cAMP means that the TSH receptor is activated and the activation of this receptor is associated with an increased

expression of the TG gene). Thus, a lower *ST6GAL1* gene expression leads to lower TSH receptor sialylation and lower TSH receptor activation. The result is a lower transcription of the TG gene. The third possibility is the association of ST6GAL1 and Tg through Tg. Tg has autoregulatory potential and can suppress its own expression [29,30]. Sue et al. suggested that Tg that is poorly iodinated or sialylated has a higher potential to trigger Tg-mediated signaling [18] and also has a higher affinity for the asialoglycoprotein (ASGP) receptor (one of the proposed receptors that could be involved in Tg-mediated signaling) [31,32]. Thus, a lower *ST6GAL1* expression could lead to a decrease in Tg sialylation. This would result in a higher concentration of poorly sialylated Tg which has a higher potential to trigger Tg-mediated signaling. Tg-mediated signaling can suppress TG gene expression. The disadvantage of this explanation is that the role of ASGPR in Tg-mediated signaling has not been thoroughly investigated, and several authors have pointed out that it is necessary to further investigate the signal transduction that occurs after Tg binding to ASGPR [31–33]. In addition, since lower *ST6GAL1* expression could result in a higher concentration of poorly sialylated Tg, this could increase the amount of Tg in the blood. Specifically, it is known that preferentially immature Tg (desialylated or poorly sialylated) is transferred to the blood by transcytosis [19,34].

In addition to the standard frequentist approach to GWA mapping, we performed a Bayesian multi-SNP mapping by fitting a BSLMM on 7,289,083 SNPs and 1096 individuals. The multi-SNP BSLMM approach uncovered additional association signals outside of the *ST6GAL1* gene. This study showed that the T allele in rs10283166 SNP located within the intronic region of the noncoding *PVT1* gene on chromosome 8 is associated with decreased plasma Tg levels. The *PVT1* gene encodes a long noncoding RNA that has an oncogenic role in various types of cancer [35]. Zhou et al. have shown that *PVT1* can contribute to tumorigenesis in thyroid cancer [36]. Additionally, Zhou et al. have shown that the silencing of *PVT1* reduces TSH receptor expression [36]. Because increased TSH receptor activation was associated with increased TG gene expression, an increase in PVT1 levels would be associated with an increase in Tg levels. Given the important role of both Tg and *PVT1* in thyroid cancer, the effect of the rs10283166 SNP on *PVT1* expression should be further investigated. On chromosome 1, the G allele, A allele, and G allele within rs78946539, rs143154928, and rs12566684 SNPs, respectively, were associated with lower plasma Tg levels. These SNPs are located within the intronic region of the *PDPN* gene. According to the GTEx portal, these SNPs affected the expression of the *RP11-474O21.5* gene in the adrenal gland, but were not associated with changes in *PDPN* gene expression. The expression of both the *RP11-474O21.5* (GEPIA database [37]) and *PDPN* [38] is increased in thyroid carcinoma. An increased expression of *PDPN* has been observed in papillary thyroid carcinoma (PTC) [38], and it has been suggested that *PDPN* may be a pro-metastatic factor in PTC [38,39]. It has been suggested that the pro-metastatic activity of *PDPN* in PTC could be through the activation of the ezrin–radixin–moesin (ERM) proteins [40]. Interestingly, moesin (ERM protein) has been shown to activate the Wnt/ $\beta$ -catenin signaling pathway [41] whose increased activation was associated with increased TG transcription (described earlier in the text) [26].

This study showed that the T allele in rs35862113 SNP located on chromosome 14 is associated with increased plasma Tg levels. This SNP is located in the intronic region of the Microtubule Affinity Regulating Kinase 3 (*MARK3*) gene. According to the GTEx portal, this SNP was associated with a reduced *MARK3* expression in thyroid tissue. Thus, lower *MARK3* expression results in increased plasma Tg levels. The possible association of *MARK3* with Tg is through Plakophilin-2 (PKP2) since PKP2 is one of the targets of *MARK3*. The phosphorylation of PKP2 by *MARK3* creates a 14–3–3 binding site [42] and it has been suggested that the phosphorylation of PKP2 by *MARK3* and subsequent binding by 14–3–3 prevents the nuclear localization of PKP2 [43]. According to Niell et al., PKP2 antagonizes Wnt/ $\beta$ -catenin signaling [44] (thus, it may consequently lead to lower TG transcription (described earlier) [26]). Additionally, this study showed that the C allele in rs1631354 SNP, located in the intronic region of retinoic acid receptor beta gene (*RARB*) on chromosome



3, is associated with increased plasma Tg levels. According to the Human Protein Atlas, RAR $\beta$  expression is high in the thyroid [13,14] while RAR $\beta$  expression is reduced in thyroid carcinomas [45,46]. One previous study showed that treatment with RAR $\beta$  binding retinoic acid (a metabolite of vitamin A) inhibited TG gene expression [47] while another showed that retinoic acid treatment increased TG gene expression [48].

Finally, allele A in rs11202702 SNP, on chromosome 10, was associated with an increase in Tg plasma levels. According to the GTEx portal, this allele is also associated with an increase in Ankyrin repeat domain-containing protein 22 (ANKRD22) expression in the esophageal mucosa (although a significant association between this SNP and ANKRD22 expression was not observed in thyroid tissue). ANKRD22 can activate the Wnt/ $\beta$ -catenin signaling pathway [49] (thus, it can consequently lead to an increase in TG transcription (described earlier in the text) [26]). This SNP (rs11202702) is located within the intronic region of the renalase gene (RNLS). To date, it has not been shown whether rs11202702 SNP affects RNLS gene expression. RNLS can activate AKT [50] which activates the Wnt/ $\beta$ -catenin signaling pathway [51] (therefore, it can consequently lead to an increase in TG transcription [26]).

In conclusion, the use of frequentist and Bayesian methods in inferring the genetic background of plasma Tg levels led to the confirmation of our previous results and the assessment of new parameters. We performed association mapping with both single-SNP and multi-SNP approaches. The results of the multi-SNP BSLMM approach are consistent with the results of our recent frequentist GWAS that showed an association of the *ST6GAL1* gene with plasma Tg levels in healthy individuals [19]. In the meta-analysis, we increased the sample size (from 1094 to 2190 healthy individuals) and with 16 confirmed variants [19], we found an additional 67 variants within the *ST6GAL1* gene associated with plasma Tg levels. We further fine-mapped the genetic architecture of Tg by estimating the PVE, PGE, and polygenic score. We found that all available variants explained approximately 17% of the variance in Tg levels and that 52% of this variation is due to a relatively small number of 16 variants that have a major effect on Tg levels. We constructed a predictive polygenic score of plasma Tg levels. Although polygenic predictions are of little use in the clinical setting, they facilitate new experimental designs and discoveries. For example, they can be used in a newly genotyped cohort to correlate the observed phenotypic traits with the genetic prediction of another trait. This approach yields a powerful design because if there exists an association between the traits, it must be due to genetic factors since there are no shared environmental factors [52]. This approach will be the scope of our future studies investigating the genetic factors underlying thyroid function. Because the most significant association signals in our meta-analysis were associated with both lower plasma Tg levels and lower *ST6GAL1* gene expression, we offered several explanations of how a lower *ST6GAL1* gene expression may lead to a decrease in plasma Tg levels. The molecular background of the influence of *ST6GAL1* on Tg levels should be examined in vitro and in vivo. Although our data strongly suggest the existence of additional effects beyond the *ST6GAL1* gene, further studies are needed to functionally characterize these complex effects. In addition, since Tg levels are altered in various thyroid diseases, the association of the identified genes in patients with different thyroid diseases needs to be examined. Moreover, our recent study observed an increase in *ST6GAL1* in various well-differentiated thyroid carcinomas (I.G., unpublished data). Finally, the conclusion of this study is that the genetic architecture of plasma Tg is not purely polygenic, but rather sparse, i.e., influenced by a few genes with major effects.

## 4. Materials and Methods

### 4.1. Study Population

This study was performed on participants originating from two Croatian cohorts: from the mainland city of Split (CROATIA\_Split) and the island of Korcula (CROATIA\_Korcula), derived from the "10,001 Dalmatians project" [53], which was part of the Croatian Biobank program. Participants were recruited from the island of Korcula in three rounds and subco-

horts were named CROATIA\_Korcula 1, CROATIA\_Korcula 2, and CROATIA\_Korcula 3, each subcohort consisting of 1000 participants. We excluded participants who could have any type of thyroid disease according to anamnestic data and detailed biochemical findings. Individuals who self-reported thyroid disorder, individuals taking thyroid medication or who underwent thyroid surgery, as well as individuals with Tg, TSH, free T3 (fT3), free T4 (fT4), Tg autoantibodies (TgAb), or thyroid peroxidase antibodies (TPOAb) levels outside of the normal reference range for our population were excluded. The published discovery phase [19] included 1094 participants from CROATIA\_Split and CROATIA\_Korcula 1 cohorts, and in the current study, we included an additional 1096 participants from the CROATIA\_Korcula 2 and CROATIA\_Korcula 3 cohorts. The final number of participants in the combined dataset for the meta-analysis was 2190. The characteristics of the cohorts are shown in Table 2. Written informed consent was obtained from participants and the study protocol was approved by the Ethical board of the University of Split, School of Medicine (No: 2181-198-03-04-14-0031 and 2181-198-03-04-19-0022).

**Table 2.** Characteristics of the study population.

Cohort	Split	Korcula 1	Korcula 2	Korcula 3
N	605	489	593	505
Women	321 (53%)	297 (61%)	328 (55.3%)	294 (58.2%)
Age	51 (39, 61)	56 (46, 67)	54 (40, 65)	54 (39, 65)
Tg	9.20 (4.80, 14.50)	10.20 (6.40, 15.70)	10.1 (5.6, 16.4)	10.6 (7.5, 16.1)

Values in the table represent median (interquartile range) or n (%). n, number of participants; Tg, thyroglobulin.

#### 4.2. Genotyping and Imputation

Genotyping platforms and quality control procedures are summarized in Supplementary Table S1. Cohorts CROATIA\_Korcula 2 and CROATIA\_Korcula 3 were genotyped together using a mix of Illumina genotyping platforms CNV370v1, CNV370-Quadv3, and OmniExpressExome-8v1-2\_A. Quality control (QC) steps were applied to all genotyping array data. The minimum call rate was 98% for SNPs and 97% for individuals, and autosomal SNPs not in Hardy–Weinberg equilibrium ( $p$ -value  $< 1 \times 10^{-6}$ ) were excluded. SHAPEIT v2.r873 and the Positional Burrows–Wheeler Transform (PBWT) [54] provided by the Wellcome Sanger Institute were used for phasing and imputing data into the Haplotype Reference Consortium (HRC) reference panel [55]. Additional QC was performed on imputed data. Imputed variants not in Hardy–Weinberg equilibrium ( $p$ -value  $< 1 \times 10^{-6}$ ), with minor allele frequency (MAF)  $< 0.01$  or with an information score  $< 0.4$ , were excluded. Sex chromosomes were not analyzed. Due to the heavy computational burden of fitting a multi-SNP approach, only variants with an information score  $\geq 0.9$  were used for the Bayesian modeling, and the compared LMM analysis. The final number of SNPs tested for association with Tg levels was 7,289,083 for both frequentist and Bayesian approaches, and 6,554,718 overlapping SNPs for the meta-analysis. Cohorts CROATIA\_Korcula 2 and CROATIA\_Korcula 3 were merged with an earlier genotyped CROATIA\_Korcula 1 cohort and this merged dataset was used for prediction analyses. The final number of SNPs used in the estimation of hyperparameters and prediction analyses was 7,289,083.

#### 4.3. Biochemical Measurements

Levels of thyroid hormones and antibodies in the plasma of participants were determined by immunoassay methods with the Liaison XL Biomedica Chemiluminescence Analyzer. Reference ranges for the study population were: Tg 0.2–50 ng/mL, TSH 0.3–3.6 mIU/L, fT3 3.39–6.47 pmol/L, fT4 10.29–21.88 pmol/L, TgAb 5–100 IU/mL, and TPOAb levels 1–16 IU/mL. All biochemical measurements were performed in the Biochemistry Laboratory in the Department of Nuclear Medicine at the University Hospital Split.

#### 4.4. Genome-Wide Association Analyses

Genome-wide association analyses in cohorts CROATIA\_Split and CROATIA\_Korcula 1 were performed in our previously conducted discovery GWAS [19]. We conducted a new GWAS in an independent combined dataset CROATIA\_Korcula2 and CROATIA\_Korcula 3 consisting of 1096 participants. For the association analysis, we considered two different approaches: the frequentist LMM and Bayesian BSLMM, both implemented using the software GEMMA 0.98.5 [56]. The phenotype used in both approaches was the same; Tg levels were firstly regressed on sex and age using R statistical software [57] and regression residuals were further quantile normalized to a standard normal distribution.

##### 4.4.1. Linear Mixed Model (LMM)

We fit a standard LMM using GEMMA 0.98.5. in the following form:

$$\mathbf{y} = \mathbf{W}\alpha + \mathbf{x}\beta + \mathbf{u} + \epsilon \quad (1)$$

$$\mathbf{u} \sim MVN_n(0, \lambda\tau^{-1}\mathbf{K}) \quad (2)$$

$$\epsilon \sim MVN_n(0, \tau^{-1}\mathbf{I}_n) \quad (3)$$

where  $\mathbf{y}$  is a vector of Tg residuals corrected for age and sex for  $n = 1096$  individuals,  $\mathbf{W}$  is a  $n \times c$  matrix of covariates (fixed effects) in our case; a column of 1s,  $\alpha$  is a  $c$ -vector of the intercept;  $\mathbf{x}$  is an  $n$ -vector of marker genotypes,  $\beta$  is the effect size of the marker,  $\mathbf{u}$  is an  $n$ -vector of random effects;  $\epsilon$  is an  $n$ -vector of errors;  $\tau^{-1}$  is the variance of the residual errors,  $\lambda$  is the ratio between the two variance components,  $\mathbf{K}$  is a known  $n \times n$  relatedness matrix and  $\mathbf{I}_n$  is an  $n \times n$  identity matrix.  $MVN_n$  denotes the  $n$ -dimensional multivariate normal distribution. Effect sizes represent the change in adjusted Tg levels for each additional effect allele in the genotypes of participants.

##### 4.4.2. Bayesian Framework

LMM implemented in GEMMA with Equation (1) tests the alternative hypothesis  $H_1 : \beta \neq 0$  against the null hypothesis  $H_0 : \beta = 0$  for each SNP in turn. Extensions of LMM that jointly account for the effects of variants across multiple loci could further increase power to detect causal variants. Bayesian LMMs are capable of modeling all markers jointly by assuming different prior distributions on the marker effects and sampling from their posterior distribution. Bayesian models developed for the estimation of the SNP effect sizes start with a simple linear model that relates genotypes  $\mathbf{X}$  to phenotypes  $\mathbf{y}$ :

$$\mathbf{y} = \mathbf{1}_n\mu + \mathbf{X}\beta + \epsilon \quad (4)$$

$$\epsilon \sim MVN_n(0, \tau^{-1}\mathbf{I}_n) \quad (5)$$

where  $\mathbf{y}$  is a vector of phenotypes measured on  $n$  individuals,  $\mathbf{X}$  is a  $n \times p$  matrix of genotypes measured on that same  $n$  individuals at  $p$  genetic markers,  $\beta$  is a  $p$ -vector of genetic marker effects,  $\mathbf{1}_n$  is an  $n$ -vector of 1s,  $\mu$  is a scalar of the phenotype mean, and  $\epsilon$  is an  $n$ -vector of error terms that have variance  $\tau^{-1}$ . Our aim was to estimate the parameter  $\beta$ , that is, the effects of genetic markers, however, since the number of genetic markers  $p$  in our study (7,289,083) was considerably larger than the number of individuals  $n$  (1096), we needed to make some modeling assumptions for SNP effect sizes  $\beta$ . These different assumptions on the priors vary from the infinitesimal (i.e., the polygenic) model which assumes that all SNPs have a non-zero effect, to the direct opposite, the sparse model which assumes that a relatively small proportion of all variants affect the phenotype. The performance of the model depends on the underlying true genetic architecture of the studied trait. However, in general, this true genetic architecture is unknown. The most commonly used polygenic modeling approach assumes that all SNPs affect the phenotype (have a non-zero effect) with normally distributed effect sizes:

$$\beta \sim N(0, \sigma_{\beta}^2) \quad (6)$$

Equation (1) with the normality assumption (6) for effect sizes  $\beta$  yields a model referred to as the linear mixed model (LMM) for its resulting random effect term of the combined genetic effects.

#### 4.4.3. Bayesian Sparse Linear Mixed Model (BSLMM)

A more general assumption, which includes both polygenic and sparse modeling assumptions as special cases, is that the effect sizes come from a mixture of two normal distributions:

$$\beta_i \sim \pi N(0, (\sigma_a^2 + \sigma_b^2)/p\tau) + (1 - \pi)N(0, \sigma_b^2/p\tau) \quad (7)$$

where  $\pi$  is the proportion of SNPs with large effects, and therefore the model is interpreted under the assumption that all variants have at least a small effect, where  $\sigma_b^2/p\tau$  is the variance of small effects, and  $\sigma_a^2/p\tau$  is the additional variance of large effects. The resulting model is the Bayesian sparse linear mixed model (BSLMM) proposed by Zhou et al. [58]. By assuming a combination of polygenic and sparse effects for the prior distribution of effect sizes, BSLMM is capable of adapting to different genetic architectures of the studied traits. Multi-SNP association mapping in BSLMM accounts for relatedness among individuals and population stratification by including a genomic kinship matrix as a random effect term. It also accounts for linkage disequilibrium (LD) between SNPs by estimating SNP effect sizes  $\beta$  while controlling for other SNPs included in the model [58]. BSLMM uses a Markov chain Monte Carlo algorithm to sample from the posterior to obtain the SNP effect size  $\beta$ . As opposed to  $p$ -values from LMM, for each SNP, it outputs a posterior inclusion probability (PIP), which is the probability that the marker is associated with the trait given the data, calculated as a proportion of chain iterations in which that SNP has a large effect. SNPs that are most robustly associated with the phenotype are therefore expected to have large PIPs and these SNPs are the most probable candidates for being the functional variants affecting plasma Tg variation. We ran a BSLMM on the same dataset (1096 individuals and 7,289,083 variants) as in our primary frequentist LMM association analysis in order to compare the single-SNP and multi-SNP approaches and to possibly reduce the incidence of false positive and false negative findings. BSLMM chain was run with default 1,000,000 sampling steps and 100,000 burn-in iterations. We used the estimated PIPs from the BSLMM output for the additional fine-mapping of the genomic regions significantly associated with Tg levels in the standard LMM analysis. The  $p$ -values from the LMM were plotted in parallel with PIPs from the BSLMM analysis in the Manhattan plots using the R package “CMplot” [59].

#### 4.5. SNP Heritability Estimation

We estimated the proportion of variance in phenotypes explained by all available genotypes (PVE), also referred to as the “chip heritability”, by assuming that the SNP effect sizes follow a mixture of two normal distributions (Equation (7)), as implemented in GEMMA BSLMM.

#### 4.6. Genetic Prediction of Thyroglobulin Levels (Polygenic Score PGS Analysis)

Predicting phenotypes from genotypes for newly observed individuals can greatly aid the development of precision medicine. However, predictions require the development of statistical methods that can accurately model the polygenic architecture of the studied trait. This is achieved by constructing a polygenic score (PGS). The simplest PGS is essentially a weighted sum of genotypes across SNPs, where weights are the estimated genetic effect sizes ( $\beta$ ) [60]. We decided to utilize the BSLMM model for genomic prediction since this method was designed for use on individual-level data and has been demonstrated to outperform several other genomic prediction methods [58]. Tg levels were firstly regressed on sex and age using the R software. Derived residuals were quantile normalized to a

standard normal distribution in R before the PGS analysis. Because GEMMA requires that the input genotype file for the PGS analysis contains both training and test data, Cohorts CROATIA\_Korcula 2 and CROATIA\_Korcula 3 were merged with the earlier genotyped CROATIA\_Korcula 1 cohort and this merged dataset was used for constructing the PGS. Sample data from the combined cohorts CROATIA\_Korcula 2 and CROATIA\_Korcula 3 were used as training data, and sample data from the CROATIA\_Korcula 1 cohort were used as test data. A Bayesian sparse linear mixed model was then fitted on the training data and its prediction performance was evaluated by calculating the Pearson's correlation coefficient between the predicted and observed values in the test data. The estimate of PVE for the SNPs used in the prediction analysis represents the potential upper bound for the performance of PGS [60]. Because of this, we expected that the prediction accuracy of the most efficient PGS would not exceed the estimated value of PVE.

#### 4.7. Meta-Analysis

We combined our previously conducted and published GWAS results in the CROATIA\_Split and CROATIA\_Korcula1 cohorts with our newly conducted GWAS in the CROATIA\_Korcula 2 and 3 cohorts using a fixed-effect inverse-variance weighted model. To visualize the meta-analysis results, a Manhattan plot and a quantile–quantile (Q-Q) plot were generated using the R package “qqman” [61]. A regional association plot for the genomic region within 500 Kb of the top hit was generated using LocusZoom software [20], and a forest plot for the most significant SNP association was generated using the R package MetABEL.

#### 4.8. GTEx Project

The Genotype-Tissue Expression (GTEx) project [23] provides the scientific community with a resource to study human gene expression and regulation and its relationship with genetic variation. By analyzing global RNA expression within individual tissues and treating the expression levels of genes as quantitative traits, variations in gene expression that are highly correlated with genetic variation can be identified as expression quantitative trait loci or eQTLs. The GTEx Project database contains the analyses of mRNA levels in 49 different tissues, including thyroid tissue obtained from 574 donors with available genotype data. The data used for the analyses described in this manuscript were obtained from the GTEx portal.

#### 4.9. Colocalization Analysis

Colocalization testing brings it closer to establishing causal relationships. If an SNP is significantly associated with both Tg levels and the gene's expression (i.e., it is an expression quantitative trait locus, eQTL), then this may suggest a regulatory role of the SNP on gene expression in the pathway to Tg levels, which can also be regarded as vertical pleiotropy. Using the LocusFocus tool [22], we tested whether our meta-analysis signals were colocalized with the eQTL signals. The LocusFocus tool implements a frequentist colocalization framework—the Simple Sum (SS) developed by Gong et al. [62]. The SS is more powerful for colocalization than existing methods, particularly in regions of high linkage disequilibrium (LD) and allelic heterogeneity. The performance of SS relative to other frequently implemented Bayesian colocalization methods designed for summary-level data was documented by Gong and collaborators [62]. To perform the colocalization analysis, we integrated our meta-analysis summary statistics data with cis-eQTL data from thyroid tissue from the GTEx project v8.

**Supplementary Materials:** The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/ijms23042173/s1>.

**Author Contributions:** Conceptualization, N.P. and T.Z.; data curation, N.P. and T.B.; formal analysis, N.P.; funding acquisition, T.Z.; investigation, I.G., T.B., V.T., A.M., A.P., O.P., and C.H.; methodology, N.P.; supervision, T.Z.; visualization, N.P.; writing—original draft, N.P., M.B.L. and T.Z.; writing—



review and editing, N.P., M.B.L., I.G. and T.Z. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work has been supported by the Croatian Science Foundation under the project “Regulation of Thyroid and Parathyroid Function and Blood Calcium Homeostasis” (No. 2593). The “10 001 Dalmatians” project was funded by grants from the Medical Research Council (UK), European Commission Framework 6 project EUROSPAN (Contract No. LSHG-CT-2006018947), the Republic of Croatia Ministry of Science, Education and Sports (grant number 216-1080315-0302), the Croatian Science Foundation (grant number 8875), CEKOM (Ministry of Economy, Entrepreneurship and Crafts), Croatian National Centre of Research Excellence in Personalized Healthcare (grant number KK.01.1.1.01.0010), and the Centre of Competence in Molecular Diagnostics (KK.01.2.2.03.0006). The work of C.H. was supported by the MRC University Unit Programme Grant MC\_UU\_00007/10 (QTL in Health and Disease).

**Institutional Review Board Statement:** This study was conducted according to the guidelines of the Declaration of Helsinki and the study protocol was approved by the Ethical board of the University of Split, School of Medicine (No: 2181-198-03-04-14-0031 and 2181-198-03-04- 19-0022).

**Informed Consent Statement:** Informed consent was obtained from all subjects involved in this study.

**Data Availability Statement:** Individual-level genetic and phenotypic data from CROATIA Split and Korcula cohorts are not available to outside researchers due to privacy restrictions. Complete summary statistics from the frequentist and Bayesian genome-wide association analyses are available.

**Acknowledgments:** We would like to thank all participants of this study and acknowledge the invaluable support of the local teams in Zagreb and Split.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

AD, Alzheimer’s disease; ANKRD22, Ankyrin repeat domain-containing protein 22; ASGP receptor, asialoglycoprotein receptor;  $\beta$ , Beta coefficients; BSLMM, Bayesian sparse linear mixed model; Chr, chromosome; fT3, free T3; fT4, free T4; eQTL, expression quantitative trait locus; ERM proteins, ezrin–radixin–moesin proteins; ETPPI, equal tail posterior probability intervals; GTEx project, Genotype-Tissue Expression project; GWA, genome-wide association; GWAS, genome-wide association study; HRC, Haplotype Reference Consortium; HWE, Hardy–Weinberg equilibrium; LD, linkage disequilibrium; LMM, linear mixed model; MAF, minor allele frequency; MARK3, Microtubule affinity regulating kinase 3; NES, normalized effect size; PBWT, Positional Burrows–Wheeler Transform; PGE, proportion of genetic variance explained by variants with major effect; PGS, polygenic score; PIP; posterior inclusion probability; PKP2, Plakophilin-2; PTC, papillary thyroid carcinoma; PVE, proportion of variance in phenotypes explained; QC, quality control; Q-Q plot, quantile–quantile plot; RARB, retinoic acid receptor beta; RNLS, renalase; SNP, single nucleotide polymorphism; SS, Simple Sum; ST6GAL1,  $\beta$ -galactoside  $\alpha$ -2,6-sialyltransferase; T3, triiodothyronine; T4, thyroxine; Tg, thyroglobulin; TgAb, Tg autoantibodies; TPOAb, thyroid peroxidase antibodies; TSH, thyroid-stimulating hormone; TTF-1, thyroid transcription factor 1.

## References

1. Citterio, C.E.; Targovnik, H.M.; Arvan, P. The role of thyroglobulin in thyroid hormonogenesis. *Nat. Rev. Endocrinol.* **2019**, *15*, 323–338. [\[CrossRef\]](#)
2. Dunn, J.T.; Dunn, A.D. The importance of thyroglobulin structure for thyroid hormone biosynthesis. *Biochimie* **1999**, *81*, 505–509. [\[CrossRef\]](#)
3. Rousset, B.; Dupuy, C.; Miot, F.; Dumont, J. Chapter 2 Thyroid Hormone Synthesis and Secretion. In *Endotext [Internet]*; Feingold, K.R., Anawalt, B., Boyce, A., Chrousos, G., de Herder, W.W., Dhatariya, K., Dungan, K., Grossman, A., Hershman, J.M., Hofland, J., et al., Eds.; MDText.com, Inc.: South Dartmouth, MA, USA, 2000.
4. Marino, M.; Chiovato, L.; Mitsiades, N.; Latrofa, F.; Andrews, D.; Tseleni-Balafouta, S.; Collins, A.B.; Pinchera, A.; McCluskey, R.T. Circulating thyroglobulin transcytosed by thyroid cells is complexed with secretory components of its endocytic receptor megalin. *J. Clin. Endocrinol. Metab.* **2000**, *85*, 3458–3467. [\[CrossRef\]](#)
5. Indrasena, B.S. Use of thyroglobulin as a tumour marker. *World J. Biol. Chem.* **2017**, *8*, 81–85. [\[CrossRef\]](#)



6. Premawardhana, L.D.; Lo, S.S.; Phillips, D.I.; Prentice, L.M.; Rees Smith, B. Variability of serum thyroglobulin levels is determined by a major gene. *Clin. Endocrinol.* **1994**, *41*, 725–729. [[CrossRef](#)]
7. Vali, M.; Rose, N.R.; Caturegli, P. Thyroglobulin as autoantigen: Structure-function relationships. *Rev. Endocr. Metab. Disord.* **2000**, *1*, 69–77. [[CrossRef](#)]
8. Mallet, B.; Lejeune, P.J.; Baudry, N.; Niccoli, P.; Carayon, P.; Franc, J.L. N-Glycans Modulate in-Vivo and in-Vitro Thyroid-Hormone Synthesis - Study at the N-Terminal Domain of Thyroglobulin. *J. Biol. Chem.* **1995**, *270*, 29881–29888. [[CrossRef](#)] [[PubMed](#)]
9. Verge, C.; Bouchatal, A.; Chirat, F.; Guerardel, Y.; Maftah, A.; Petit, J.M. Involvement of ST6Gal I-mediated alpha2,6 sialylation in myoblast proliferation and differentiation. *FEBS Open Bio* **2020**, *10*, 56–69. [[CrossRef](#)] [[PubMed](#)]
10. Grollman, E.F.; Doi, S.Q.; Weiss, P.; Ashwell, G.; Wajchenberg, B.L.; Medeiros-Neto, G. Hyposialylated thyroglobulin in a patient with congenital goiter and hypothyroidism. *J. Clin. Endocrinol. Metab.* **1992**, *74*, 43–48.
11. Grollman, E.F.; Saji, M.; Shimura, Y.; Lau, J.T.; Ashwell, G. Thyrotropin regulation of sialic acid expression in rat thyroid cells. *J. Biol. Chem.* **1993**, *268*, 3604–3609. [[CrossRef](#)]
12. Kiljanski, J.; Ambroziak, M.; Pachucki, J.; Jazdzewski, K.; Wiechno, W.; Stachlewska, E.; Gornicka, B.; Bogdanska, M.; Nauman, J.; Bartoszewicz, Z. Thyroid sialyltransferase mRNA level and activity are increased in Graves' disease. *Thyroid* **2005**, *15*, 645–652. [[CrossRef](#)] [[PubMed](#)]
13. Thul, P.J.; Akesson, L.; Wiking, M.; Mahdessian, D.; Geladaki, A.; Ait Blal, H.; Alm, T.; Asplund, A.; Bjork, L.; Breckels, L.M.; et al. A subcellular map of the human proteome. *Science* **2017**, *356*, eaal3321. [[CrossRef](#)] [[PubMed](#)]
14. Uhlen, M.; Fagerberg, L.; Hallstrom, B.M.; Lindskog, C.; Oksvold, P.; Mardinoglu, A.; Sivertsson, A.; Kampf, C.; Sjostedt, E.; Asplund, A.; et al. Tissue-based map of the human proteome. *Science* **2015**, *347*, 1260419. [[CrossRef](#)] [[PubMed](#)]
15. Hedlund, M.; Ng, E.; Varki, A.; Varki, N.M. Alpha 2-6-linked sialic acids on N-glycans modulate carcinoma differentiation in vivo. *Cancer Res.* **2008**, *68*, 388–394. [[CrossRef](#)]
16. Fenouillet, E.; Fayet, G.; Hovsepian, S.; Bahraoui, E.M.; Ronin, C. Immunochemical Evidence for a Role of Complex Carbohydrate Chains in Thyroglobulin Antigenicity. *J. Biol. Chem.* **1986**, *261*, 15153–15158. [[CrossRef](#)]
17. Salabe, H.; Dominici, R.; Salabe, G.B. Immunological Properties of Tg Carbohydrates—Enhancement of Tg Immunoreaction by Removal of Sialic-Acid. *Clin. Exp. Immunol.* **1976**, *25*, 234–243.
18. Sue, M.; Hayashi, M.; Kawashima, A.; Akama, T.; Tanigawa, K.; Yoshihara, A.; Hara, T.; Ishido, Y.; Ito, T.; Takahashi, S.; et al. Thyroglobulin (Tg) activates MAPK pathway to induce thyroid cell growth in the absence of TSH, insulin and serum. *Biochem. Biophys. Res. Commun.* **2012**, *420*, 611–615. [[CrossRef](#)]
19. Matana, A.; Popović, M.; Boutin, T.; Torlak, V.; Brdar, D.; Gunjača, I.; Kolčić, I.; Boraska Perica, V.; Punda, A.; Rudan, I.; et al. Genetic Variants in the ST6GAL1 Gene Are Associated with Thyroglobulin Plasma Level in Healthy Individuals. *Thyroid* **2019**, *29*, 886–893. [[CrossRef](#)]
20. Pruim, R.J.; Welch, R.P.; Sanna, S.; Teslovich, T.M.; Chines, P.S.; Gliedt, T.P.; Boehnke, M.; Abecasis, G.R.; Willer, C.J. LocusZoom: Regional visualization of genome-wide association scan results. *Bioinformatics* **2010**, *26*, 2336–2337. [[CrossRef](#)]
21. Altshuler, D.M.; Durbin, R.M.; Abecasis, G.R.; Bentley, D.R.; Chakravarti, A.; Clark, A.G.; Donnelly, P.; Eichler, E.E.; Flicek, P.; Gabriel, S.B.; et al. An integrated map of genetic variation from 1092 human genomes. *Nature* **2012**, *491*, 56–65.
22. Panjwani, N.; Wang, F.; Mastromatteo, S.; Bao, A.; Wang, C.; He, G.; Gong, J.; Rommens, J.M.; Sun, L.; Strug, L.J. LocusFocus: Web-based colocalization for the annotation and functional follow-up of GWAS. *PLoS Comput. Biol.* **2020**, *16*, e1008336. [[CrossRef](#)] [[PubMed](#)]
23. Lonsdale, J.; Thomas, J.; Salvatore, M.; Phillips, R.; Lo, E.; Shad, S.; Hasz, R.; Walters, G.; Garcia, F.; Young, N.; et al. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **2013**, *45*, 580–585. [[CrossRef](#)] [[PubMed](#)]
24. Wei, A.; Fan, B.; Zhao, Y.; Zhang, H.; Wang, L.; Yu, X.; Yuan, Q.; Yang, D.; Wang, S. ST6Gal-I overexpression facilitates prostate cancer progression via the PI3K/Akt/GSK-3beta/beta-catenin signaling pathway. *Oncotarget* **2016**, *7*, 65374–65388. [[CrossRef](#)] [[PubMed](#)]
25. Gilbert-Sirieix, M.; Makoukji, J.; Kimura, S.; Talbot, M.; Caillou, B.; Massaad, C.; Massaad-Massade, L. Wnt/beta-catenin signaling pathway is a direct enhancer of thyroid transcription factor-1 in human papillary thyroid carcinoma cells. *PLoS ONE* **2011**, *6*, e22280. [[CrossRef](#)]
26. Civitareale, D.; Lonigro, R.; Sinclair, A.J.; Di Lauro, R. A thyroid-specific nuclear protein essential for tissue-specific expression of the thyroglobulin promoter. *EMBO J.* **1989**, *8*, 2537–2542. [[CrossRef](#)] [[PubMed](#)]
27. Frenzel, R.; Krohn, K.; Eszlinger, M.; Tonjes, A.; Paschke, R. Sialylation of human thyrotropin receptor improves and prolongs its cell-surface expression. *Mol. Pharmacol.* **2005**, *68*, 1106–1113. [[CrossRef](#)] [[PubMed](#)]
28. Korta, P.; Pochec, E. Glycosylation of thyroid-stimulating hormone receptor. *Endokrynol. Pol.* **2019**, *70*, 86–100. [[CrossRef](#)]
29. Huang, H.; Shi, Y.; Liang, B.; Cai, H.; Cai, Q. Iodinated TG in Thyroid Follicular Lumen Regulates TTF-1 and PAX8 Expression via TSH/TSHR Signaling Pathway. *J. Cell Biochem.* **2017**, *118*, 3444–3451. [[CrossRef](#)]
30. Suzuki, K.; Lavaroni, S.; Mori, A.; Ohta, M.; Saito, J.; Pietrarello, M.; Singer, D.S.; Kimura, S.; Katoh, R.; Kawaoi, A.; et al. Autoregulation of thyroid-specific gene transcription by thyroglobulin. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 8251–8256. [[CrossRef](#)]
31. Sellitti, D.F.; Suzuki, K. Intrinsic Regulation of Thyroid Function by Thyroglobulin. *Thyroid* **2014**, *24*, 625–638. [[CrossRef](#)]
32. Ulianich, L.; Suzuki, K.; Mori, A.; Nakazato, M.; Pietrarello, M.; Goldsmith, P.; Pacifico, F.; Consiglio, E.; Formisano, S.; Kohn, L.D. Follicular thyroglobulin (TG) suppression of thyroid-restricted genes involves the apical membrane asialoglycoprotein receptor and TG phosphorylation. *J. Biol. Chem.* **1999**, *274*, 25099–25107. [[CrossRef](#)] [[PubMed](#)]

33. Luo, Y.; Ishido, Y.; Hiroi, N.; Ishii, N.; Suzuki, K. The Emerging Roles of Thyroglobulin. *Adv. Endocrinol.* **2014**, *2014*, 1–7. [[CrossRef](#)]
34. Marino, M.; McCluskey, R.T. Role of thyroglobulin endocytic pathways in the control of thyroid hormone release. *Am. J. Physiol. Cell Physiol.* **2000**, *279*, C1295–C1306. [[CrossRef](#)] [[PubMed](#)]
35. Onagoruwa, O.T.; Pal, G.; Ochu, C.; Ogunwobi, O.O. Oncogenic Role of PVT1 and Therapeutic Implications. *Front. Oncol.* **2020**, *10*, 17. [[CrossRef](#)] [[PubMed](#)]
36. Zhou, Q.; Chen, J.; Feng, J.; Wang, J. Long noncoding RNA PVT1 modulates thyroid cancer cell proliferation by recruiting EZH2 and regulating thyroid-stimulating hormone receptor (TSHR). *Tumor Biol.* **2016**, *37*, 3105–3113. [[CrossRef](#)] [[PubMed](#)]
37. Tang, Z.; Li, C.; Kang, B.; Gao, G.; Li, C.; Zhang, Z. GEPIA: A web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res.* **2017**, *45*, W98–W102. [[CrossRef](#)] [[PubMed](#)]
38. Rudzińska, M.; Gaweł, D.; Sikorska, J.; Karpińska, K.M.; Kiedrowski, M.; Stępień, T.; Marchlewska, M.; Czarnocka, B. The Role of Podoplanin in the Biology of Differentiated Thyroid Cancers. *PLoS ONE* **2014**, *9*, e96541. [[CrossRef](#)]
39. Tseng, C.P.; Leong, K.K.; Liou, M.J.; Hsu, H.L.; Lin, H.C.; Chen, Y.A.; Lin, J.D. Circulating epithelial cell counts for monitoring the therapeutic outcome of patients with papillary thyroid carcinoma. *Oncotarget* **2017**, *8*, 77453–77464. [[CrossRef](#)]
40. Sikorska, J.; Gaweł, D.; Domek, H.; Rudzińska, M.; Czarnocka, B. Podoplanin (PDPN) affects the invasiveness of thyroid carcinoma cells by inducing ezrin, radixin and moesin (E/R/M) phosphorylation in association with matrix metalloproteinases. *BMC Cancer* **2019**, *19*, 85. [[CrossRef](#)]
41. Zhu, X.; Morales, F.C.; Agarwal, N.K.; Dogruluk, T.; Gagea, M.; Georgescu, M.M. Moesin Is a Glioma Progression Marker That Induces Proliferation and Wnt/ $\beta$ -Catenin Pathway Activation via Interaction with CD44. *Cancer Res.* **2013**, *73*, 1142–1155. [[CrossRef](#)]
42. Müller, J.; Ritt, D.A.; Copeland, T.D.; Morrison, D.K. Functional analysis of C-TAK1 substrate binding and identification of PKP2 as a new C-TAK1 substrate. *EMBO J.* **2003**, *22*, 4431–4442. [[CrossRef](#)] [[PubMed](#)]
43. Hatzfeld, M.; Wolf, A.; Keil, R. Plakophilins in Desmosomal Adhesion and Signaling. *Cell Commun. Adhes.* **2014**, *21*, 25–42. [[CrossRef](#)] [[PubMed](#)]
44. Niell, N.; Larriba, M.J.; Ferrer-Mayorga, G.; Sánchez-Pérez, I.; Cantero, R.; Real, F.X.; Del Peso, L.; Muñoz, A.; González-Sancho, J.M. The human PKP2/plakophilin-2 gene is induced by Wnt/ $\beta$ -catenin in normal and colon cancer-associated fibroblasts. *Int. J. Cancer* **2018**, *142*, 792–804. [[CrossRef](#)] [[PubMed](#)]
45. Hoftijzer, H.C.; Liu, Y.Y.; Morreau, H.; van Wezel, T.; Pereira, A.M.; Corssmit, E.P.M.; Romijn, J.A.; Smit, J.W.A. Retinoic acid receptor and retinoid X receptor subtype expression for the differential diagnosis of thyroid neoplasms. *Eur. J. Endocrinol.* **2009**, *160*, 631–638. [[CrossRef](#)] [[PubMed](#)]
46. Czajka, A.A.; Wójcicka, A.; Kubiak, A.; Kotlarek, M.; Bakuła-Zalewska, E.; Koperski, L.; Wiechno, W.; Jażdżewski, K. Family of microRNA-146 Regulates RAR $\beta$  in Papillary Thyroid Carcinoma. *PLoS ONE* **2016**, *11*, e0151968. [[CrossRef](#)] [[PubMed](#)]
47. Namba, H.; Yamashita, S.; Morita, S.; Villadolid, M.C.; Kimura, H.; Yokoyama, N.; Izumi, M.; Ishikawa, N.; Ito, K.; Nagataki, S. Retinoic acid inhibits human thyroid peroxidase and thyroglobulin gene expression in cultured human thyrocytes. *J. Endocrinol. Invest.* **1993**, *16*, 87–93. [[CrossRef](#)]
48. Kurebayashi, J.; Tanaka, K.; Otsuki, T.; Moriya, T.; Kunisue, H.; Uno, M.; Sonoo, H. All-Trans-Retinoic Acid Modulates Expression Levels of Thyroglobulin and Cytokines in a New Human Poorly Differentiated Papillary Thyroid Carcinoma Cell Line, KTC-11. *J. Clin. Endocrinol. Metab.* **2000**, *85*, 2889–2896.
49. Wu, Y.; Liu, H.; Gong, Y.; Zhang, B.; Chen, W. ANKRD22 enhances breast cancer cell malignancy by activating the Wnt/ $\beta$ -catenin pathway via modulating NuSAP1 expression. *Bosn. J. Basic Med. Sci.* **2021**, *21*, 294–304. [[CrossRef](#)]
50. Pointer, T.C.; Gorelick, F.S.; Desir, G.V. Renalase: A Multi-Functional Signaling Molecule with Roles in Gastrointestinal Disease. *Cells* **2021**, *10*, 2006. [[CrossRef](#)]
51. Thompson, M.; Nejak-Bowen, K.; Monga, S.P.S. Crosstalk of the Wnt Signaling Pathway. In *Targeting the Wnt Pathway in Cancer*; Goss, K.H., Kahn, M., Eds.; Springer: New York, NY, USA, 2011; pp. 51–80.
52. Visscher, P.M.; Wray, N.R.; Zhang, Q.; Sklar, P.; McCarthy, M.I.; Brown, M.A.; Yang, J. 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **2017**, *101*, 5–22. [[CrossRef](#)]
53. Rudan, I.; Marusic, A.; Jankovic, S.; Rotim, K.; Boban, M.; Lauc, G.; Grkovic, I.; Dogas, Z.; Zemunik, T.; Vatavuk, Z.; et al. “10001 Dalmatians”: Croatia launches its national biobank. *Croat. Med. J.* **2009**, *50*, 4–6. [[CrossRef](#)] [[PubMed](#)]
54. Durbin, R. Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT). *Bioinformatics* **2014**, *30*, 1266–1272. [[CrossRef](#)] [[PubMed](#)]
55. McCarthy, S.; Das, S.; Kretzschmar, W.; Delaneau, O.; Wood, A.R.; Teumer, A.; Kang, H.M.; Fuchsberger, C.; Danecek, P.; Sharp, K.; et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **2016**, *48*, 1279–1283. [[PubMed](#)]
56. Zhou, X.; Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **2012**, *44*, 821–824. [[CrossRef](#)] [[PubMed](#)]
57. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016.
58. Zhou, X.; Carbonetto, P.; Stephens, M. Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genet.* **2013**, *9*, e1003264. [[CrossRef](#)] [[PubMed](#)]

- 
59. Yin, L.; Zhang, H.; Tang, Z.; Xu, J.; Yin, D.; Zhang, Z.; Yuan, X.; Zhu, M.; Zhao, S.; Li, X.; et al. rMVP: A Memory-efficient, Visualization-enhanced, and Parallel-accelerated tool for Genome-Wide Association Study. *Genomics Proteom. Bioinform.* **2021**, *in press*. [[CrossRef](#)] [[PubMed](#)]
  60. Ma, Y.; Zhou, X. Genetic prediction of complex traits with polygenic scores: A statistical review. *Trends Genet.* **2021**, *37*, 995–1011. [[CrossRef](#)]
  61. Turner, S.D. qqman: An R package for visualizing GWAS results using Q-Q and manhattan plots. *J. Open Source Softw.* **2018**, *3*, 731. [[CrossRef](#)]
  62. Gong, J.F.; Wang, F.; Xiao, B.W.; Panjwani, N.; Lin, F.; Keenan, K.; Avolio, J.; Esmaili, M.; Zhang, L.; He, G.M.; et al. Genetic association and transcriptome integration identify contributing genes and tissues at cystic fibrosis modifier loci. *PLoS Genet.* **2019**, *15*, e1008007. [[CrossRef](#)]