Journal of *Personalized Medicine*

MDPI

Article

Development of a Machine Learning Model to Predict Non-Durable Response to Anti-TNF Therapy in Crohn's Disease Using Transcriptome Imputed from Genotypes

Soo Kyung Park [1,2,†], Yea Bean Kim [3,†], Sangsoo Kim [3,*], Chil Woo Lee [2], Chang Hwan Choi [4], Sang-Bum Kang [5], Tae Oh Kim [6], Ki Bae Bang [7], Jaeyoung Chun [8], Jae Myung Cha [9], Jong Pil Im [10], Min Suk Kim [11], Kwang Sung Ahn [12], Seon-Young Kim [13] and Dong Il Park [1,2,*]

1 Division of Gastroenterology, Department of Internal Medicine and Inflammatory Bowel Disease Center, Kangbuk Samsung Hospital, School of Medicine, Sungkyunkwan University, Seoul 03181, Korea; skparkmd@gmail.com
2 Medical Research Institute, Kangbuk Samsung Hospital, School of Medicine, Sungkyunkwan University, Seoul 03181, Korea; chilwoo.lee@gmail.com
3 Department of Bioinformatics, Soongsil University, Seoul 06978, Korea; yebins96@gmail.com
4 Department of Internal Medicine, College of Medicine, Chung-Ang University, Seoul 06973, Korea; gicch@cau.ac.kr
5 Department of Internal Medicine, Daejeon St. Mary's Hospital, College of Medicine, Catholic University, Daejeon 34943, Korea; sangucsd@gmail.com
6 Department of Internal Medicine, Haeundae Paik Hospital, College of Medicine, Inje University, Busan 48108, Korea; kto0440@paik.ac.kr
7 Department of Internal Medicine, College of Medicine, Dankook University, Cheonan 31116, Korea; kibaebang@gmail.com
8 Department of Internal Medicine, Gangnam Severance Hospital, College of Medicine, Yonsei University, Seoul 06273, Korea; j40479@gmail.com
9 Department of Internal Medicine, Kyung Hee University Hospital at Gang Dong, College of Medicine, Kyung Hee University, Seoul 05278, Korea; clicknox@hanmail.net
10 Department of Internal Medicine and Liver Research Institute, College of Medicine, Seoul National University, Seoul 03080, Korea; jpim0911@snu.ac.kr
11 Department of Human Intelligence and Robot Engineering, Sangmyung University, Cheonan 31066, Korea; minsuk.kim@smu.ac.kr
12 Functional Genome Institute, PDXen Biosystems Inc., Suwon 16488, Korea; kwangsung.ahn@gmail.com
13 Personalized Medicine Research Center, Korea Research Institute of Bioscience and Biotechnology (KRIBB), Daejeon 34141, Korea; kimsy@kribb.re.kr
* Correspondence: sskimb@ssu.ac.kr (S.K.); diksmc.park@samsung.com (D.I.P.); Tel.: +82-2-820-0457 (S.K.); +82-2-2001-8555 (D.I.P.)
† These authors contributed equally to this work.

Abstract: Almost half of patients show no primary or secondary response to monoclonal anti-tumor necrosis factor α (anti-TNF) antibody treatment for inflammatory bowel disease (IBD). Thus, the exact mechanisms of a non-durable response (NDR) remain inadequately defined. We used our genome-wide genotype data to impute expression values as features in training machine learning models to predict a NDR. Blood samples from various IBD cohorts were used for genotyping with the Korea Biobank Array. A total of 234 patients with Crohn's disease (CD) who received their first anti-TNF therapy were enrolled. The expression profiles of 6294 genes in whole-blood tissue imputed from the genotype data were combined with clinical parameters to train a logistic model to predict the NDR. The top two and three most significant features were genetic features (*DPY19L3*, *GSTT1*, and *NUCB1*), not clinical features. The logistic regression of the NDR vs. DR status in our cohort by the imputed expression levels showed that the β coefficients were positive for *DPY19L3* and *GSTT1*, and negative for *NUCB1*, concordant with the known eQTL information. Machine learning models using imputed gene expression features effectively predicted NDR to anti-TNF agents in patients with CD.

Keywords: genotype; genetic features; anti-TNF; Crohn's disease

check for updates

Citation: Park, S.K.; Kim, Y.B.; Kim, S.; Lee, C.W.; Choi, C.H.; Kang, S.-B.; Kim, T.O.; Bang, K.B.; Chun, J.; Cha, J.M.; et al. Development of a Machine Learning Model to Predict Non-Durable Response to Anti-TNF Therapy in Crohn's Disease Using Transcriptome Imputed from Genotypes. *J. Pers. Med.* **2022**, *12*, 947. https://doi.org/10.3390/jpm12060947

Academic Editors: Aristotelis Chatziioannou and Yudong Zhang

Received: 19 April 2022
Accepted: 8 June 2022
Published: 9 June 2022

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

1. Introduction

Crohn's disease (CD) is a chronic relapsing inflammatory bowel disease (IBD) that causes progressive bowel damage and disability [1]. Monoclonal anti-tumor necrosis factor α (anti-TNF) antibodies have revolutionized the care of patients with CD, enabling the achievement of clinical and endoscopic remission [2]. However, despite the established efficacy of these drugs, one-fifth of patients will not respond to these agents (primary non-response (PNR)), and an additional one-third will eventually fail therapy (secondary loss of response, non-durable response (NDR)), requiring an additional or changed medication or surgery [3–5]. The exact mechanisms of PNR and NDR remain poorly defined. There is currently considerable interest and an unmet need for the use of genetic markers to predict therapeutic responses. Most prior studies examining this question studied one or a few candidate genes, had small sample sizes, and did not yield definitive results [6–14]. Polymorphisms in TNF-α [8], IBD5 locus [15], immunoglobulin G (IgG) Fc receptor IIIa [16], autophagy (ATG16L1) [17], and apoptosis-related genes [6] have also been variably associated with a response to anti-TNF agents. A limitation of exclusively studying a few candidate loci or IBD-risk alleles is the possibility of missing potentially relevant associations across loci that more broadly influence immune function across a disease spectrum.

It is tempting to develop machine learning models for the prediction of the durable response (DR) status of anti-TNF therapy based on patient genotypes, as they are non-invasive and baseline in nature. However, it is challenging because there are numerous patient genotype markers to consider, and the number of cases is limited. As far as we know, there have been no reports on genotype-based markers that can predict future NDR in CD. The straightforward use of genome-wide genotype data for model training is usually hampered by overfitting (data not shown). Hence, here we propose a novel approach to dimensionality reduction by transforming the genotype dataset into a gene expression dataset (Figure 1).

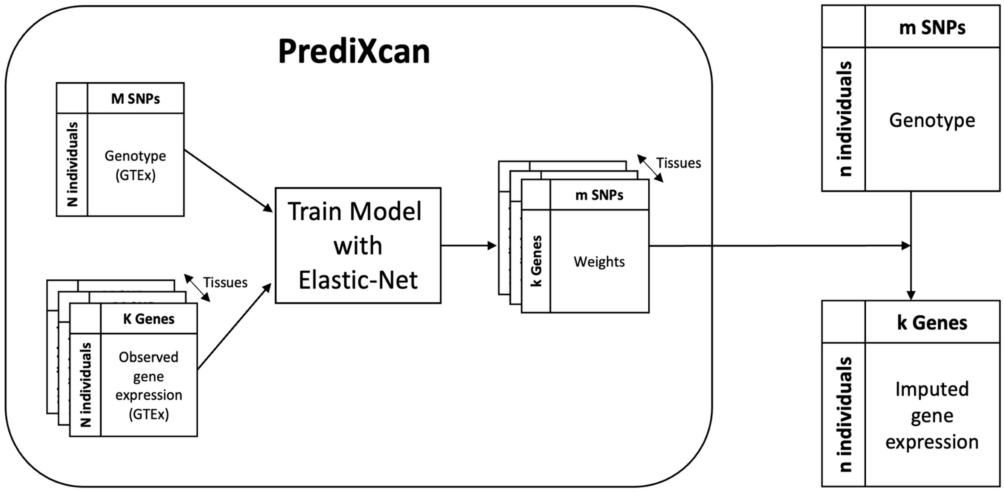

Figure 1. Overview of imputing gene expression from genotype. The PrediXcan models (round-cornered box) were downloaded from https://predictdb.org (accessed on 21 April 2021). They were developed for a number of tissues using the matched genotype and gene expression datasets compiled by the GTEx consortium. A given gene's expression value was linearly modeled from the genotypes of neighboring single-nucleotide proteins (SNPs), which was selected through elastic net. By applying the linear regression weights to our genotype data (upper right), we were able to impute the gene expression in our tissue of interest (lower right).

A transcriptome-wide association study that tests the association between the phenotype and the "imputed" gene expression has been successfully applied in many cases. The PrediXcan is one such method that imputes gene expression from genotypes using

machine learning models developed based on the GTEx genotype and the corresponding transcriptome datasets of various tissues. In this study, we applied the PrediXcan to our genome-wide genotype data to impute gene expression and used the resulting expression values as well as clinical parameters as features for training machine learning models for the prediction of NDR vs. DR status. Thus, the number of features was reduced from millions to thousands. Notably, imputed expression values are not direct experimental observations; rather, they are derived features upon the combination of experimental values in a mathematically defined manner.

2. Materials and Methods

2.1. Study Population

A total of 894 IBD patients were recruited from the Identification of the Mechanism of the occurrence and Progression of Crohn's disease through integrated Analysis on both genetiC and environmenTal factors (IMPACT), UC multiomics, and Occurrence of Anti-drug antibody and change of drug level after CT-P13 therapy and their Impact on clinical outcomes in moderate to Severe inflammatory bowel disease (OACIS) cohorts. IMPACT was a prospective multicenter study established in Korea in 2017. Clinical data and biological specimens (including blood, stool, and tissue specimens) of CD patients who were newly diagnosed or followed up within 16 university hospitals were collected. The ulcerative colitis (UC) multiomics study was a prospective multicenter study, established in Korea in 2020. A total of 14 university hospitals participated in this study and collected clinical data and biological specimens (including blood, stool, tissue, and saliva specimens) from UC patients. Details of the IMPACT and UC multiomics study cohorts have been previously described [18]. The OACIS study was a prospective multicenter observational study conducted at 18 university hospitals in Korea between August 2016 and September 2019. Consecutive patients older than 18 years with moderate to severe active CD or UC who started CT-P13 therapy were prospectively enrolled in the study.

Among the 894 patients, the inclusion criteria were as follows: (1) diagnosis of CD and (2) first anti-TNF therapy consisting of infliximab or CT-P13. Using a chart review, two study investigators (S.K.P., D.I.P.) characterized the patients' responses to their first anti-TNF therapy. All patients received standard induction dosing (infliximab or CT-P13 5 mg/kg at weeks 0, 2, and 6, and every 8 weeks thereafter). DR was defined as the maintenance of the response to anti-TNF therapy for at least 24 months after initiation. NDR was defined as non-response within 24 months after starting therapy accompanied by an alteration in therapy (addition or escalation of corticosteroids, switch to a different agent, or surgery). Patients with primary non-response (non-response at 12 weeks after starting therapy), those who ceased treatment due to adverse effects before the 24-month time point, and those who experienced adverse events related to a loss of response (for example, infusion reactions due to immunogenicity) were classified as NDR.

2.2. Genotyping

Blood samples from the three cohorts were used for genotyping with the Korea Biobank Array, which comprises 833,535 single-nucleotide proteins (SNPs) [19]. Sample quality control of call rate (>95%), heterozygosity (within ± 3 standard deviation of mean), relative relationship (proportion of IBD < 0.2), and principal component analysis (within the main cluster) tests eliminated 16 samples, resulting in 878 samples. All QC metrics were calculated with PLINK (v1.90b6.4). The genotype data of 749,383 SNPs that survived the quality control analysis were imputed with the Korean reference panel using the BEAGLE software package [20]. The reference panel comprised 28,445 samples that were genotyped with the Korea Biobank Array and subsequently imputed with the East Asian population of the 1000 Genomes data. After removal of the SNPs with minor allele frequency < 0.05 or that violated the Hardy–Weinberg equilibrium ($p < 0.001$), a total of 6,153,437 SNPs were available for analysis.

Among the 878 samples that passed filtering, 234 patients met the inclusion criteria (220 and 14 with DR and NDR, respectively). Table 1 shows the patients' demographic and epidemiological characteristics. For expression quantitative trait locus (eQTL) analysis of the association signals, GTEx [21] and the CD eQTL database from the Asan Medical Center IBD eQTL Browser (http://asan.crohneqtl.com/ accessed on 30 October 2021) were utilized [22].

Table 1. Patients' clinical characteristics.

	Non-Durable Response (n = 14)	Durable Response (n = 220)	*p*-Value
Age at diagnosis, year (SD)	26.3 (9.2)	28.2 (9.1)	0.45
Gender, male (%)	8 (57.1%)	162 (73.6%)	0.21
History of smoking, n (%)	5 (35.7%)	47 (21.4%)	0.20
Family history of IBD, n (%)	0 (0%)	7 (3.2%)	1.0
Disease duration, year (SD)	9.1 (5.5)	7.5 (3.8)	0.16
Disease location, n (%)			0.07
Ileal	7 (50%)	52 (23.6%)	
Colonic	2 (14.3%)	31 (14.1%)	
Ileocolonic	5 (35.7%)	137 (62.3%)	
Upper GI involvement, n (%)	0 (0%)	11 (5.0%)	0.39
Disease behavior, n (%)			0.35
Inflammatory	9 (64.3%)	163 (74.1%)	
Stricturing	1 (7.1%)	25 (11.4%)	
Penetrating	4 (28.6%)	32 (14.5%)	
Perianal disease, n (%)	6 (42.9%)	85 (38.6%)	0.75
Combination immunosuppressants, n (%)	10 (71.4%)	206 (93.6%)	<0.001
Intestinal resection, n (%)	6 (42.9%)	61 (27.7%)	0.22

IBD, inflammatory bowel disease; GI, gastrointestinal.

2.3. Imputing Gene Expression from Genotype

Machine learning models for imputing tissue-specific gene expression from genotypes were downloaded from the PrediXcan website (https://predictdb.org/post/2017/11/29/gtex-v7-expression-models/ accessed on 21 April 2021). Each was a linear regression model that was trained on the European subset of the genotype and expression datasets compiled by GTEx v7. The PrediXcan algorithm considers only SNPs proximal to the target gene and further limits them by employing a feature selection method based on an elastic net. The imputed expression value of the target gene is then obtained by taking the scalar product of the regression coefficients with the alternative allele counts of the surviving SNPs. It should be noted that the imputability of a gene varies across tissues, and PrediX can evaluate imputability through cross-validation, releasing only the models passing preset filtering criteria. Among the tissue models that PrediXcan v7 released, we considered that gene models for the whole blood, transverse colon, and small intestine terminal ileum were relevant to CD, focusing the subsequent analyses on these tissues only. We calculated the imputed expression values of 6294, 5612, and 3107 genes for 234 patients in those 3 tissues, respectively. The imputed expression values were used as features in the predictive model development, as described below. The selected genes were also tested for their associations with NDR/DR status using the univariate logistic regression implemented as a PrediXcan function.

2.4. Development of Predictive Machine Learning Models

2.4.1. Preprocessing

The following clinical variables—age, sex, smoking status, family history, disease location at diagnosis (including upper gastrointestinal involvement), disease behavior at diagnosis (including perianal involvement), and gene expression values imputed by the PrediXcan—were used as the features for the predictive model for discriminating between NDR (cases) and DR (controls). Several functions in scikit-learn [23] have been used to encode and standardize variables. For example, numerical variables, such as age and expression values, were standardized using StanardScaler. Dichotomous categorical variables were encoded as 0 or 1 using LabelEncoder, whereas multiclass features were binarized using OneHotEncoder.

2.4.2. Model Training and Feature Selection

A total of 234 samples were split into training and test sets in an 8:2 ratio using the train_test_split of scikit-learn, while constraining the 14 cases to be split into 10 and 4 in the training and test sets, respectively. For the model training, we used logistic regression of scikit-learn, as it is simple and sufficiently effective for large cycles of iterative training with an additional option of feature selection, such as Least Absolute Shrinkage and Selection Operator (LASSO) and elastic net. Our feature selection process proceeded in two stages. In the first stage, we leveraged the LASSO penalty in the logistic regression training to reduce the number of features. Specifically, the C parameter in logistic regression was scanned from 10 to 100 in increments of 10, and the trained model was evaluated by 5-fold cross-validation of the area under the receiver operating characteristic curve (AUC-ROC). Using the liblinear optimizer in the logistic regression, 10–1000 features survived. The overall scheme of the model training is shown in Figure 2.

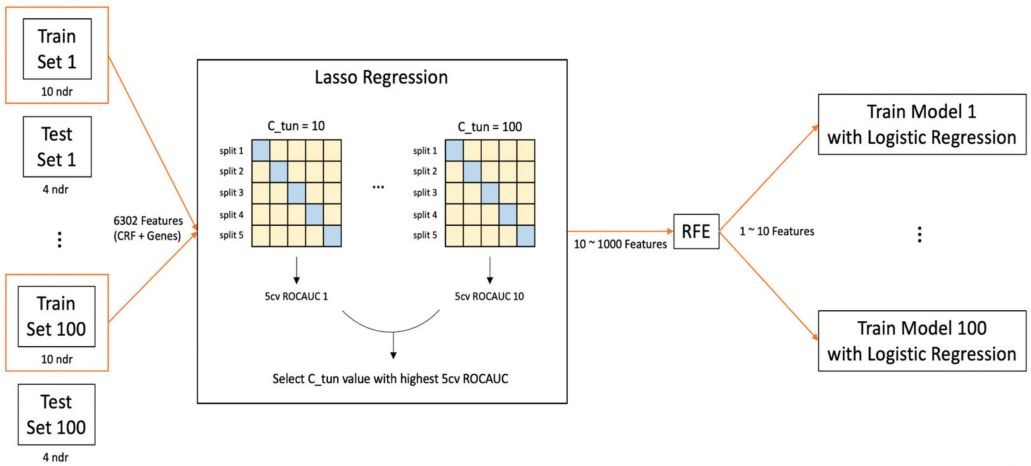

Figure 2. The overall model training scheme. The dataset was split into training and test sets in an 8:2 ratio. This random split was repeated 100 times. For each split, the model training involving Least Absolute Shrinkage and Selection Operator (LASSO) regression, and recursive feature elimination (RFE) was performed. In the LASSO regression, the C parameter was scanned from 10 to 100 in multiples of 10 for the highest 5-fold cross-validation (5-CV) area under the receiver operating characteristic curve (AUC-ROC) value. For the best C parameter, typically 10 to 1000 features survived. Among these features, a fixed number, ranging from 1 to 10, of features was selected through RFE. For a given number of the selected features, 100 different models were developed due to the 100 random data splits. The model performance was evaluated with the test set using logistic regression of the selected feature(s). For 5-CV, the training and test sets are shown in light yellow and blue, respectively.

In the second stage of the feature selection process, we used recursive feature elimination (RFE) of scikit-learn, which iterates model training by removing the least important

feature in each round until the desired number of features survive. Using the selected features, the model performance was measured using the test set. We repeated the entire feature selection stage 100 times with random shuffling of the training and test split. The most frequent feature combinations were sought from this repetition.

As our data are somewhat imbalanced and limited in size, there is a concern that AUC-ROC can discern different models. For each model, we also evaluated the precision-recall curve and AUC-PRC. Given an ROC curve, the Youden's index, sensitivity + specificity, was calculated, and the cutoff was defined where the Youden's index was maximal. With this cutoff, sensitivity (recall), specificity, and precision of the model were assessed. The mean and standard deviation of these metrics from the 100 repetitions were then reported. Among the 100 repetitions, the case whose metrics were the most similar to the mean values was chosen and its ROC and PRC curves were presented as representative.

3. Results

3.1. DR/NDR Prediction Models Based on Different Sampling Tissues

GTEx collected the genotype data and corresponding expression datasets from various sampling tissues. Accordingly, the PrediXcan provides a gene expression imputation model, one for each gene in each sampled tissue. The number of imputable genes varies across tissues. As we were interested in predicting the NDR to infliximab in IBD patients, we downloaded the imputable gene models on three sampling tissues: whole blood (6294 genes), colon transverse (5612 genes), and small intestine terminal ileum (3107 genes). First, we imputed the gene expression values from our genotype dataset by using these models for each tissue. Second, the imputed expression values were used as features in training NDR/DR response models based on logistic regression with the feature selection scheme based on the LASSO penalty and RFE. The performance of the model at each selection step was evaluated using 5-fold cross-validation of the AUC-ROC. The test performance was measured for the best model. Since our dataset was limited to a small number of NDR samples, our results may not be stable and are likely to suffer from overfitting. To overcome this limitation, we repeated the entire feature selection process 100 times and searched for a recurring combination of features.

As the first step of the proof of concept, we selected only one feature from the feature selection processes involving the LASSO and RFE. Figure 3 shows the distribution of the predictive performance of these models during 100 repetitions in three different sampling tissues. Although the performance with the test set varied approximately twice as much as that with the training set, the medians were very similar to each other, implying that the extent of overfitting was negligible. See Figure S1 for the distributions of AUC-PRC and Figure S2 for representative ROC and PRC curves.

Among the three different tissue models, the imputed expression in whole blood achieved the highest performance, and the difference in performance among the models from different tissues implies that genetic factors contribute significantly to the predictability of NDR/DR status. Indeed, the features that were selected most frequently were genes in all three tissue models (Table 2 and Table S1).

Table 2. Most frequently selected single feature for each tissue expression imputation model and its performance in classifying NDR vs. DR.

Tissue Expression Model	Selected Feature	Selection Frequency	AUC-ROC (SD)	
			Training 5-CV Set	Test Set
Whole blood	*DPY19L3*	79/100	0.845 (0.027)	0.839 (0.070)
Colon transverse	*TXNDC16*	40/100	0.728 (0.060)	0.711 (0.150)
Small intestine terminal ileum	*ENSG00000270127*	14/100	0.738 (0.050)	0.720 (0.120)

5CV, five-fold cross-validation; AUC-ROC, area under the receiver operating characteristic curve; SD, standard deviation; NDR, non-durable response; DR, durable response.

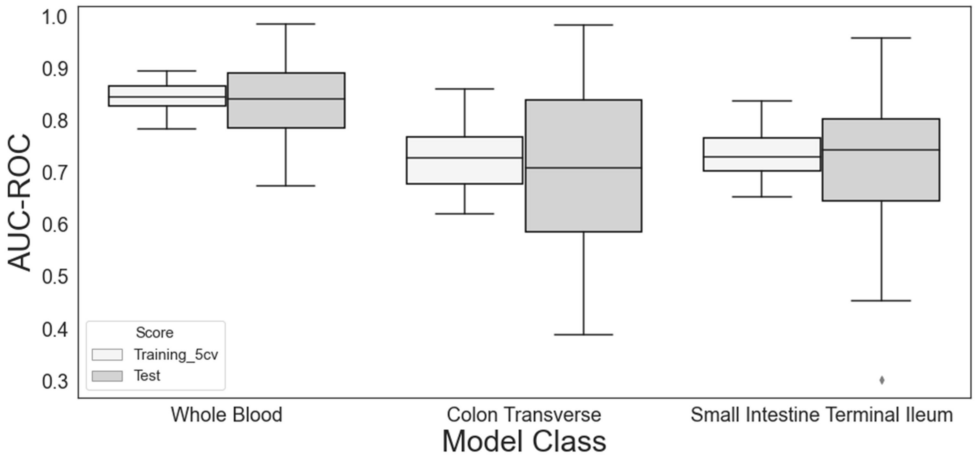

Figure 3. Performance of the three tissue models used for gene expression imputation in classifying the non-durable response (NDR) vs. the durable response (DR). The model training by 5-fold cross-validation and feature selection via LASSO and recursive feature elimination was repeated 100 times (see main text for details). The single most significant gene was selected from each trial. The training and test performances given as area under the receiver operating characteristic curve (AUC-ROC) are shown as boxplots. That from the whole-blood model was significantly higher than that from the colon transverse ($p_{t\text{-test}} = 6.5 \times 10^{-38}$ and 3.7×10^{-12} for the training and test sets, respectively) or the small intestine terminal ileum ($p_{t\text{-test}} = 1.8 \times 10^{-46}$ and 7.5×10^{-15} for the training and test sets, respectively).

Interestingly, Dpy-19-like C-mannosyltransferase 3 (*DPY19L3*) was selected highly consistently (79 of 100 trials) from the random shuffling of the training/test dataset split in the whole-blood model, which performed the best. On the other hand, two different genes, *TXNDC16* and *ENSG00000270127*, were most frequently selected in the other two tissue models. Furthermore, their selections were less consistent than those of the whole-blood model. Based on these observations, we focused on the whole-blood model in the following analyses.

3.2. Selection of Top Two and Three Features Using Whole-Blood Model

We selected the top two and three most significant features by combining the LASSO penalty and the RFE. These selections were independently performed for each of the 100 random shufflings. As shown in Table 3, only the genetic features were selected. *DPY19L3*, the single most frequently selected feature, was included in the most frequent combinations of two and three features. Glutathione S-transferase theta 1 (*GSTT1*) was the most frequent two- and three-feature selection feature.

Table 3. The most frequently selected combination of two or three genes for the whole-blood expression imputation model and its performance for classifying NDR vs. DR.

No. of Features	Selected Feature Set	Selection Frequency	AUC-ROC (SD)	
			Training 5CV Set	Test Set
1	*DPY19L3*	79	0.845 (0.027)	0.839 (0.070)
2	*DPY19L3, GSTT1*	32	0.918 (0.023)	0.919 (0.040)
3	*DPY19L3, GSTT1, NUCB1*	9	0.935 (0.024)	0.935 (0.040)

5CV, five-fold cross-validation; AUC-ROC, area under the receiver operating characteristic curve; SD, standard deviation; NDR, non-durable response; DR, durable response.

As shown in Figures S3 and S4 and Table S2, more features were included and better performance was achieved, although the difference between the two- and three-feature models was not large.

3.3. Contribution of Clinical Features

Although eight clinical features were included in the training, none survived the feature selection process when we selected up to three features. To evaluate the contribution of clinical features in predicting NDR/DR status, we built a logistic regression model based solely on them and compared it with the models built by infusing the genetic features selected in the previous section. Using the same scheme of random shuffling of the training/test split, the performances were evaluated (Table S3). The model with only the clinical features did not perform as well as the single gene model in terms of AUC-ROC (*DPY19L3*; 0.568 vs. 0.845 for the training set, 0.603 vs. 0.839 for the test set). In contrast, infusing genes into the model with only the clinical features dramatically improved its performance (Figure 4). This trend is similar to that observed in gene-only models. Although the inclusion of clinical features did not show clearly improved performance over the gene-only models in terms of AUC-ROC or AUC-PRC (Figures S5 and S6), the average sensitivity, specificity, and precision were improved with the inclusion of the clinical features (Table S4).

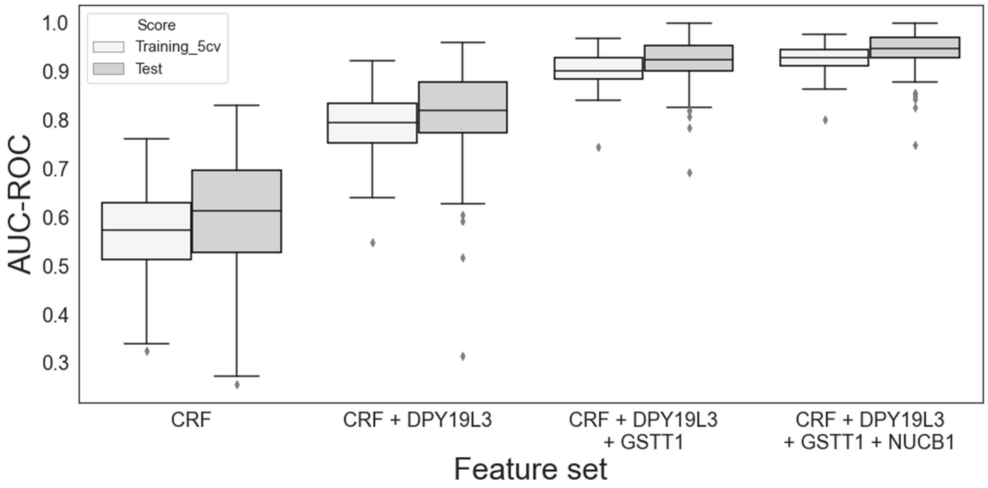

Figure 4. Performance evaluation of the models built with varying sets of features for classifying the non-durable response (NDR) vs. the durable response (DR). The feature genes were the same as listed in Table 3. A total of eight patient clinical parameters (see Methods Section), denoted as *CRF* in the figure, were also included in the model without further feature selections.

3.4. Genetic Bases of Selected Features

PrediXcan develops models for the imputation of gene expression based on the genotypes and transcriptomics datasets compiled by GTEx [24], and as most of the whole-blood samples collected by GTEx were of European origin [25], there might be concern about whether the gene expression models from PrediXcan are valid for the current study population of Korean origin. To address this, we examined the genetic basis of the three genes identified in this study. The association of the imputed expression values with NDR/DR status was analyzed by univariate logistic regression for each gene using the PrediXcan function. As shown in Table 4, all showed significant associations ($p < 0.01$). The logistic regression coefficients (β) of *DPY19L3* and *GSTT1* were positive, meaning that their higher imputed expression levels increase the probability of being NDR, while that of *NUCB1* was negative, decreasing the probability of being NDR.

Table 4. Univariate logistic regression analysis of the association between gene expression and NDR/DR status.

Gene Name	Chr	*p*-Value	β Value
DPY19L3	19	0.000965	2.703
GSTT1	22	0.00343	1.735
NUCB1	19	0.00684	−2.142

Chr, chromosome; NDR, non-durable response; DR, durable response.

For *DPY19L3*, many eQTLs are known in whole blood according to the GTEx; for example, the alternative allele of rs4805759 has a net effect size of −0.23 ($p = 5.2 \times 10^{-38}$). The eQTL database constructed for the whole-blood samples of Korean CD patients [22] also shows the same direction of effect, that is, this alternative allele has the regression slope of −0.478 ($p = 9.9 \times 10^{-6}$). In our genotype data, the alternative allele frequencies were 0.8341 and 0.4286 for DR and NDR samples, respectively ($p = 2.2 \times 10^{-5}$). The finding that the NDR samples have fewer alternative alleles of rs4805759 than DR and its eQTL regression coefficient is negative is in agreement with the univariate logistic regression coefficient (β) of *DPY19L3* being positive (Table 4). For *GSTT1*, the Korean eQTL database lists three SNPs as eQTLs (rs368588, rs2236620, and rs2236621, with slopes of 0.683–1.186), whereas no eQTLs were identified in GTEx. While *NUCB1* is not listed in the Korean eQTL database, its eQTLs are listed in the GTEx (max. net effect size of −0.45 for rs10415881). Their directions of effect were also in agreement with the respective univariate logistic coefficients (β).

4. Discussion

A recent trend in large-scale association studies is transcriptome-wide association studies that transform genome-wide genotype datasets into imputed gene expression datasets to identify gene–trait associations. The PrediXcan is one such method that imputes tissue-specific gene expression from genotypes by the machine learning models that have been pre-developed based on GTEx datasets. In this study, we applied the PrediXcan to our genome-wide genotype data and used the resulting expression values in whole-blood tissue as well as clinical parameters as features in training machine learning models for predicting NDR vs. DR status. The selected top two and three most significant features were genetic features only (*DPY19L3*, *GSTT1*, and *NUCB1*). Adding clinical features without further feature selection showed slight improvement in the performance over the gene-only models. The logistic regression of the NDR vs. DR status by the imputed expression levels of *DPY19L3, GSTT1,* and *NUCB1* were consistent with the known eQTL information retrieved from GTEx and the Korean CD eQTL database [19]. It was reported that these eQTL databases had concordant directions of eQTL for more than 96% of the target genes [19]. This supports the validity of the PrediXcan models based on the European data applied to the Korean cases.

Among the genetic features, *DPY19L3* was the most frequently selected. *DPY19L3*-mediated C-mannosylation of R-spondin1 at W^{156} is required for R-spondin1 secretion [26]. R-spondin1 is a secreted protein that enhances Wnt signaling, an important pathway for immune cell maintenance and renewal [26]. Wnt signaling in immune cells is very diverse, for example, the tolerogenic role of dendritic cells, development of natural killer cells, thymopoiesis of T cells, B-cell-driven initiation of T cells, and macrophage actions in tissue repair, regeneration, and fibrosis [27]. As the imputed expression level of *DPY19L3* is supposed to be higher in NDR than in DR, the inflammatory response of immune cells might not be controlled by an anti-TNF agent.

GSTT1 was the most frequently used two-feature selection method. Among the members of GSTs, glutathione S-transferase theta 1 (*GSTT1*) and GSTM1 in particular have become recent targets of active investigation into their role in increased susceptibility to IBD, Behçet's disease, or other autoimmune diseases such as primary sclerosing cholangitis.

GSTT1 contributes to detoxifying chemicals, including reactive oxygen species (ROS) [28]. In a previous study using a DSS-induced colitis mouse model [28], the authors noted attenuation of colitis through gene transfer of *Gstt1* via an IL-22-dependent pathway. Downregulation of *GSTT1* by the pathogen-associated molecular patterns of microbes reduces innate defense responses and goblet cell differentiation. *GSTT1* ameliorates colitis, and its mutations are linked to chronic intestinal inflammation due to insufficient dimerization.

Impaired ROS production due to inactivation of patient variants in genes encoding nicotinamide adenine dinucleotide phosphate oxidases as ROS sources is associated with CD and pancolitis, whereas ROS overproduction due to upregulation of oxidases or altered mitochondrial function has been linked to ileitis and ulcerative colitis [29]. The major role of TNF is to regulate the immune system through the activation of TNF receptors and downstream pathways involving molecules, such as nuclear factor kappa-B, mitogen-activated protein kinases, caspases, and ROS/reactive nitrogen species [30]. As the imputed expression level of *DPY19L3* is supposed to be higher in NDR than in DR, it is hypothesized that ROS levels uncontrolled by anti-TNF agents may be related to a continuous inflammatory response.

NUCB1, which is involved in three-feature selection, interacts with *GNAI1* or *GNAI3* to activate them [31]. In a mouse model, *GNAI1* and *GNAI 3* suppressed DSS-plus-azoxymethane-induced colon tumor development, whereas the expression of *GNAI2* in $CD11c^+$ cells and interleukin-6 (IL6) in $CD4^+/CD11b^+$ dendritic cells appeared to promote these effects. As GNAI1 and GNAI3 block IL6 signaling to inhibit inflammation or tumorigenesis, strategies to induce GNAI1 or block GNAI2 and IL6 have been suggested to prevent or treat colitis-associated cancer [32]. As the imputed expression level of *NUCB1* is supposed to be lower in patients with NDR, the actions of GNAI1 or GNAI3 and blocking IL6 signaling may be impaired. Despite anti-TNF agent use, the inflammatory response by IL6 might result in NDR.

The strength of our study is that we found genetic features using a machine learning method with genome-wide genotype data and the resulting expression values. Although we used the PrediXcan model based on the European origin genotypes and transcriptomic datasets compiled by GTEx, it was valid for the current Korean population. Since the gene expression value used as a feature is imputed by an individual's unique genotype, it can accordingly be viewed as an expected baseline gene expression value, rather than being related to disease onset or drug administration.

There are also some limitations to this study. First, our definitions of NDR were based on clinical evidence from chart reviews rather than from prospectively collected data. Future studies should prospectively include comprehensive clinical, endoscopic, and radiological evidence to define the response. Second, imputed expression values are not direct experimental observations; rather, they are derived features by the combination of experimental values in a mathematically defined manner. Third, except for eight clinical variables, drug history, such as combinations of immunosuppressants, was not considered in this analysis. Including more clinical variables and comparing with genetic features in the predictive model are needed in future studies.

5. Conclusions

In conclusion, machine learning models with transcriptomes imputed from genome-wide genotype datasets effectively predicted NDR to anti-TNF agents in patients with CD. However, the genetic features derived from our study require validation in another cohort, whereas the pathway of genetic features associated with NDR to anti-TNF agents requires further study.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/jpm12060947/s1, Figure S1: Performance of the models with various tissue models of PrediXcan, selecting the top significant feature for classifying a non-durable response vs. a durable response. Figure S2: Representative ROC and PRC curves of the test set for various tissue models. Figure S3: Performance of the models with increasing number of selected genetic features for

classifying a non-durable response vs. a durable response. Figure S4: Representative ROC and PRC curves of the test set for increasing number of selected features. Figure S5: Performance of the models with increasing number of genetic features and fixed contribution of clinical features for classifying a non-durable response vs. a durable response. Figure S6: Representative ROC and PRC curves of the test set for the selected genetic features with a fixed contribution of clinical features. Table S1: Average performance of various tissue models with the top significant feature. Table S2: Average performance of the whole-blood models without clinical features. Table S3: Average performance of the whole-blood models with fixed clinical features. Table S4: Average sensitivity, specificity, and precision of the whole-blood models with the test set.

Author Contributions: Conceptualization, D.I.P. and S.K.; methodology, S.K.; software, Y.B.K.; formal analysis, Y.B.K. and S.K.; investigation, S.K.P.; data curation, C.W.L., C.H.C., S.-B.K., T.O.K., K.B.B., J.C., J.M.C., J.P.I. and M.S.K.; writing—original draft preparation, S.K.P. and Y.B.K.; writing—review and editing, S.-Y.K., D.I.P. and S.K.; supervision, K.S.A.; project administration, D.I.P.; funding acquisition, D.I.P. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by a National Research Foundation (NRF) grant funded by the Korean government (NRF 2020R1A2B5B02002259).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board (or Ethics Committee) of Kangbuk Samsung Hospital and each center (protocol code: KBSMC 2016-07-029, date of approval: 18 October 2016).

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

Data Availability Statement: The genotype data generated and analyzed in the current study are not publicly available due to the limitation of study consent regarding repository deposition, but they are available from the corresponding author upon reasonable request. The imputed gene expression datasets with the clinical phenotype information are available from the zenodo database (https://doi.org/10.5281/zenodo.6464129, accessed on 16 April 2022).

Acknowledgments: The Korean Reference genotype dataset was released by the National Biobank of Korea, Korea National Institute of Health, Osong, Korea (nih.go.kr/biobank, accessed on 18 July 2019), under accession number 2019-032.

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

References

1. Cosnes, J.; Gower–Rousseau, C.; Seksik, P.; Cortot, A. Epidemiology and natural history of inflammatory bowel diseases. *Gastroenterology* **2011**, *140*, 1785–1794. [CrossRef] [PubMed]
2. D'Haens, G.R.; Panaccione, R.; Higgins, P.D.; Vermeire, S.; Gassull, M.; Chowers, Y.; Hanauer, S.B.; Herfarth, H.; Hommes, D.W.; Kamm, M.; et al. The London Position Statement of the World Congress of Gastroenterology on Biological Therapy for IBD with the European Crohn's and Colitis Organization: When to start, when to stop, which drug to choose, and how to predict response? *Am. J. Gastroenterol.* **2011**, *106*, 199–212. [CrossRef] [PubMed]
3. Allez, M.; Karmiris, K.; Louis, E.; Van Assche, G.; Ben-Horin, S.; Klein, A.; Van der Woude, J.; Baert, F.; Eliakim, R.; Katsanos, K.; et al. Report of the ECCO pathogenesis workshop on anti-TNF therapy failures in inflammatory bowel diseases: Definitions, frequency and pharmacological aspects. *J. Crohn's Colitis* **2010**, *4*, 355–366. [CrossRef] [PubMed]
4. Chowers, Y.; Sturm, A.; Sans, M.; Papadakis, K.; Gazouli, M.; Harbord, M.; Jahnel, J.; Mantzaris, G.J.; Meier, J.; Mottet, C.; et al. Report of the ECCO workshop on anti-TNF therapy failures in inflammatory bowel diseases: Biological roles and effects of TNF and TNF antagonists. *J. Crohn's Colitis* **2010**, *4*, 367–376. [CrossRef]
5. Gisbert, J.P.; Panés, J. Loss of response and requirement of infliximab dose intensification in Crohn's disease: A review. *Am. J. Gastroenterol.* **2009**, *104*, 760–767. [CrossRef]
6. Hlavaty, T.; Pierik, M.; Henckaerts, L.; Ferrante, M.; Joossens, S.; Van Schuerbeek, N.; Noman, M.; Rutgeerts, P.; Vermeire, S. Polymorphisms in apoptosis genes predict response to infliximab therapy in luminal and fistulizing Crohn's disease. *Aliment. Pharmacol. Ther.* **2005**, *22*, 613–626. [CrossRef]
7. Jürgens, M.; Laubender, R.P.; Hartl, F.; Weidinger, M.; Seiderer, J.; Wagner, J.; Wetzke, M.; Beigel, F.; Pfennig, S.; Stallhofer, J.; et al. Disease activity, ANCA, and IL23R genotype status determine early response to infliximab in patients with ulcerative colitis. *Am. J. Gastroenterol.* **2010**, *105*, 1811–1819. [CrossRef]

8. Mascheretti, S.; Hampe, J.; Kühbacher, T.; Herfarth, H.; Krawczak, M.; Fölsch, U.R.; Schreiber, S. Pharmacogenetic investigation of the TNF/TNF-receptor system in patients with chronic active Crohn's disease treated with infliximab. *Pharm. J.* **2002**, *2*, 127–136. [CrossRef]
9. Siegel, C.A.; Melmed, G.Y. Predicting response to Anti-TNF Agents for the treatment of crohn's disease. *Ther. Adv. Gastroenterol.* **2009**, *2*, 245–251. [CrossRef]
10. Taylor, K.D.; Yang, H.; Landers, C.J.; Rotter, J.I.; Targan, S.R.; Plevy, S.E.; Barry, M.J. ANCA pattern and LTA haplotype relationship to clinical responses to anti-TNF antibody treatment in Crohn's disease. *Gastroenterology* **2001**, *120*, 1347–1355. [CrossRef]
11. Vermeire, S.; Louis, E.; Rutgeerts, P.; De Vos, M.; Van Gossum, A.; Belaiche, J.; Pescatore, P.; Fiasse, R.; Pelckmans, P.; Vlietinck, R.; et al. NOD2/CARD15 does not influence response to infliximab in Crohn's disease. *Gastroenterology* **2002**, *123*, 106–111. [CrossRef]
12. Verstockt, B.; Verstockt, S.; Dehairs, J.; Ballet, V.; Blevi, H.; Wollants, W.J.; Breynaert, C.; Van Assche, G.; Vermeire, S.; Ferrante, M. Low TREM1 expression in whole blood predicts anti-TNF response in inflammatory bowel disease. *EBioMedicine* **2019**, *40*, 733–742. [CrossRef]
13. Verstockt, B.; Verstockt, S.; Blevi, H.; Cleynen, I.; de Bruyn, M.; Van Assche, G.; Vermeire, S.; Ferrante, M. TREM-1, the ideal predictive biomarker for endoscopic healing in anti-TNF-treated Crohn's disease patients? *Gut* **2019**, *68*, 1531–1533. [CrossRef]
14. Gaujoux, R.; Starosvetsky, E.; Maimon, N.; Vallania, F.; Bar-Yoseph, H.; Pressman, S.; Weisshof, R.; Goren, I.; Rabinowitz, K.; Waterman, M.; et al. Cell-centred meta-analysis reveals baseline predictors of anti-TNFα non-response in biopsy and blood of patients with IBD. *Gut* **2019**, *68*, 604–614. [CrossRef]
15. Urcelay, E.; Mendoza, J.L.; Martínez, A.; Fernández, L.; Taxonera, C.; Díaz-Rubio, M.; de la Concha, E.G. IBD5 polymorphisms in inflammatory bowel disease: Association with response to infliximab. *World J. Gastroenterol.* **2005**, *11*, 1187–1192. [CrossRef]
16. Louis, E.; El Ghoul, Z.; Vermeire, S.; Dall'Ozzo, S.; Rutgeerts, P.; Paintaud, G.; Belaiche, J.; De Vos, M.; Van Gossum, A.; Colombel, J.F.; et al. Association between polymorphism in IgG Fc receptor IIIa coding gene and biological response to infliximab in Crohn's disease. *Aliment. Pharmacol. Ther.* **2004**, *19*, 511–519. [CrossRef]
17. Koder, S.; Repnik, K.; Ferkolj, I.; Pernat, C.; Skok, P.; Weersma, R.K.; Potočnik, U. Genetic polymorphism in ATG16L1 gene influences the response to adalimumab in Crohn's disease patients. *Pharmacogenomics* **2015**, *16*, 191–204. [CrossRef]
18. Park, S.K.; Kim, S.; Lee, G.Y.; Kim, S.Y.; Kim, W.; Lee, C.W.; Park, J.L.; Choi, C.H.; Kang, S.B.; Kim, T.O.; et al. Development of a Machine Learning Model to Distinguish between Ulcerative Colitis and Crohn's Disease Using RNA Sequencing Data. *Diagnostics* **2021**, *11*, 2365. [CrossRef]
19. Moon, S.; Kim, Y.J.; Han, S.; Hwang, M.Y.; Shin, D.M.; Park, M.Y.; Lu, Y.; Yoon, K.; Jang, H.M.; Kim, Y.K.; et al. The Korea Biobank Array: Design and Identification of Coding Variants Associated with Blood Biochemical Traits. *Sci. Rep.* **2019**, *9*, 1382. [CrossRef]
20. Browning, B.L.; Zhou, Y.; Browning, S.R. A One-Penny Imputed Genome from Next-Generation Reference Panels. *Am. J. Hum. Genet.* **2018**, *103*, 338–348. [CrossRef]
21. 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]
22. Jung, S.; Liu, W.; Baek, J.; Moon, J.W.; Ye, B.D.; Lee, H.S.; Park, S.H.; Yang, S.K.; Han, B.; Liu, J.; et al. Expression Quantitative Trait Loci (eQTL) Mapping in Korean Patients with Crohn's Disease and Identification of Potential Causal Genes Through Integration With Disease Associations. *Front. Genet.* **2020**, *11*, 486. [CrossRef] [PubMed]
23. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
24. Gamazon, E.R.; Wheeler, H.E.; Shah, K.P.; Mozaffari, S.V.; Aquino-Michaels, K.; Carroll, R.J.; Eyler, A.E.; Denny, J.C.; Nicolae, D.L.; Cox, N.J.; et al. A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **2015**, *47*, 1091–1098. [CrossRef] [PubMed]
25. Mikhaylova, A.V.; Thornton, T.A. Accuracy of Gene Expression Prediction From Genotype Data With PrediXcan Varies Across and Within Continental Populations. *Front. Genet.* **2019**, *10*, 261. [CrossRef]
26. Niwa, Y.; Suzuki, T.; Dohmae, N.; Simizu, S. Identification of DPY19L3 as the C-mannosyltransferase of R-spondin1 in human cells. *Mol. Biol. Cell* **2016**, *27*, 744–756. [CrossRef]
27. Haseeb, M.; Pirzada, R.H.; Ain, Q.U.; Choi, S. Wnt Signaling in the Regulation of Immune Cell and Cancer Therapeutics. *Cells* **2019**, *8*, 1380. [CrossRef]
28. Kim, J.H.; Ahn, J.B.; Kim, D.H.; Kim, S.; Ma, H.W.; Che, X.; Seo, D.H.; Kim, T.I.; Kim, W.H.; Cheon, J.H.; et al. Glutathione S-transferase theta 1 protects against colitis through goblet cell differentiation via interleukin-22. *FASEB J.* **2020**, *34*, 3289–3304. [CrossRef]
29. Aviello, G.; Knaus, U.G. ROS in gastrointestinal inflammation: Rescue or Sabotage? *Br. J. Pharmacol.* **2017**, *174*, 1704–1718. [CrossRef]
30. Blaser, H.; Dostert, C.; Mak, T.W.; Brenner, D. TNF and ROS Crosstalk in Inflammation. *Trends Cell Biol.* **2016**, *26*, 249–261. [CrossRef]
31. Rebhan, M.; Chalifa-Caspi, V.; Prilusky, J.; Lancet, D. GeneCards: Integrating information about genes, proteins and diseases. *Trends Genet.* **1997**, *13*, 163. [CrossRef]
32. Li, Z.W.; Sun, B.; Gong, T.; Guo, S.; Zhang, J.; Wang, J.; Sugawara, A.; Jiang, M.; Yan, J.; Gurary, A.; et al. GNAI1 and GNAI3 Reduce Colitis-Associated Tumorigenesis in Mice by Blocking IL6 Signaling and Down-regulating Expression of GNAI2. *Gastroenterology* **2019**, *156*, 2297–2312. [CrossRef]