TRABAJO FIN DE MASTER EN BIOESTADÍSTICA

PREDICTIVE ANALYSIS TO FIND GERMLINE GENETIC SUSCEPTIBILITY ASSOCIATED WITH THE TUMORAL IMMUNE INFILTRATION IN PANCREATIC CANCER

JULIO 2021

Laura Gutiérrez García

Silvia Pineda San Juan, Teresa Pérez Pérez, Núria Malats Riera

INDEX

LIST OF TABLES

LIST OF FIGURES

SUMMARY

The immune system plays an important role in the tumor microenvironment since there is an interaction between tumor cells and immune cells that affects the tumor development. In particular, in pancreatic cancer, it has been studied that after characterizing B and T cell repertoire, patients have shown a large heterogeneity among them. Additionally, it was previously demonstrated that genetic susceptibility may explain around 40% of the immune system differences across individuals.

Thus, in this project, the main objective was to predict tumoral immune infiltration in pancreatic cancer patients using germline genetic variants (SNPs). T and B cell receptors were extracted from RNAseq data in 120 individuals with pancreatic cancer and richness and diversity were assessed using Expression and Entropy measures. Then, four machine learning methods were proposed (Elastic Net, Ridge Regression, Random Forest and Neural Network) focus on dealing with high dimensionality and multicollinearity problems present in high-throughput data.

The performance of the four different methods was assessed through Pearson correlation. Predictions obtained by these methods were benchmarked across 10 testing subsets in three different scenarios. Neural Network which showed the highest and the most consistent correlations between observed and predicted values, overcomes the overfitting and over-specificity problems. Being able to predict the immune infiltration with genetic variants will allow us to integrate and decipher new biological insights extremely necessary in pancreatic cancer research.

Keywords: Immune system, tumor microenvironment, machine learning, *Neural Network*, high-dimensionality, pancreatic cancer.

RESUMEN

El sistema inmunológico desempeña un papel fundamental en el microentorno del tumor, ya que, existe una interacción entre las células tumorales y las inmunes influyendo en su desarrollo. En particular, en cáncer de páncreas. Previamente, se ha estudiado que tras caracterizar el repertorio de las células B y T, los pacientes han mostrado una gran heterogeneidad entre ellos. Además, se ha demostrado que la susceptibilidad genética puede explicar hasta un 40% de las diferencias inmunes observadas entre individuos.

Así, en este trabajo, se plantea el objetivo de predecir la infiltración tumoral inmune en individuos con cáncer de páncreas usando variantes genéticas en línea germinal (SNPs). Los receptores de las células B y T se extrajeron de RNAseq de 120 individuos con cáncer de páncreas y la riqueza y diversidad se midieron mediante las medidas de Expresión y Entropía. Se proponen entonces cuatro métodos de *machine learning* (*Elastic Net*, *Ridge Regression*, *Random Forest* y *Neural Network*) enfocados a lidiar con los problemas de alta dimensionalidad y multicolinealidad presentes en nuestros datos.

La actuación de los cuatro métodos se evaluó a través de la correlación de Pearson. Las predicciones obtenidas por estos métodos fueron comparadas a lo largo de 10 subconjuntos de *testing* en tres escenarios diferentes. *Neural Network,* el cual mostró las correlaciones más altas y consistentes entre los valores predichos y observados, superó los problemas de sobreajuste y sobre-especificidad. Ser capaz de predecir la infiltración inmunológica mediante variantes genéticas nos permitirá integrar y descifrar nuevo conocimiento muy necesario para avanzar en el cáncer de páncreas.

Palabras clave: Sistema inmune, microentorno del tumor*, machine learning*, *Neural Network*, alta dimensionalidad, cáncer de páncreas.

1. INTRODUCTION

1.1 Pancreatic cancer

Pancreatic cancer (PC) is a dreadful disease usually diagnosed at an advanced stage and, despite its relatively low population incidence, it is the deadliest cancer worldwide with a 7%-5 year survival rate[1]. Important attempts have been done to advance in deciphering the complexity of PC etiology at both genetic and non-genetic risk factors (Figure 1). Regarding the non-genetic factors: obesity[2], tobacco[3], and type 2 diabetes (T2D)[4], heavy alcohol consumption[5], chronic pancreatitis[6] and ABO blood group[7] are stablished risk factors for PC. Family history[8] has been also associated with increased risk of PC while nasal allergies and asthma[9] have been associated with a reduce risk of PC. On the other hand, although relatively few Single Nucleotide Polymorphism (SNPs) associates with PC have been identified, novel candidate variants, which are located in genes with an important role in the function of pancreatic acinar cell, have been discovered to be involved in this complex disease[10]. Interestingly, the majority of known risk factors point to a chronic inflammatory process and different forms of inflammation play critical roles in tumor development which might result in immunological infiltration within the tumor.

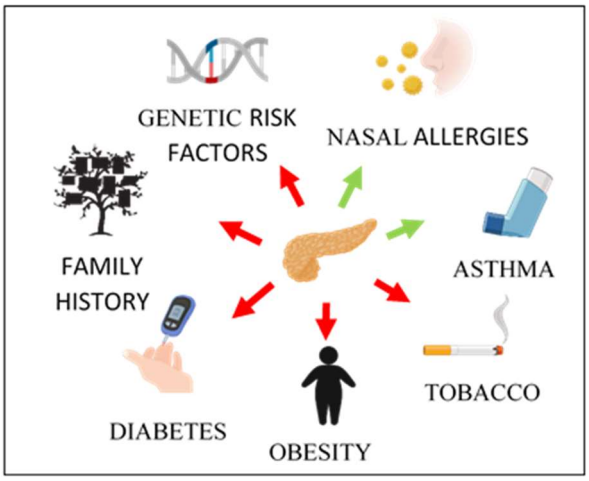

Figure 1. Some of the main risk (red arrows) and protective factors (green arrows) studied in pancreatic cancer. Image created with Biorender.com

Despite all these advances, PC is a very complex and heterogeneous disease and there is still a lack of information to characterize both genetic and non-genetic risk factors participating in its etiology. Further research is needed, specially focusing on revealing the heterogeneity of the immunological infiltration which can play an important role in the development of novel therapeutic targets[11].

1.2 Immunological infiltration

Tumors are composed of different cell populations and, among them, a wide variety of non-cancerous cell types. In almost all tumors, it exists an interaction between malignant cells and leukocyte infiltration which includes innate immune cells (tumor associated macrophages (TAMs), mast cells, Natural Killer (NK) cells, and NK T cells), and adaptive immune cells (T and B lymphocytes) (Figure 2) and a variety of stromal cells, among other components. This is known as tumor microenvironment, which currently represents an important topic in the cancer research field[12–15].

The tumor microenvironment is complex and it is well known that different characteristics such as carcinogenic pathways, mutations and the clinicopathological factors interact with the adaptive immune system. The adaptive immune system is composed of B and T lymphocytes which produce B cell receptors (BCR) or antibodies capable of recognize foreign substance, such as pathogens or viruses and T cell receptors (TCR) which recognize fragments of antigens presented on the surface of the cells[16].

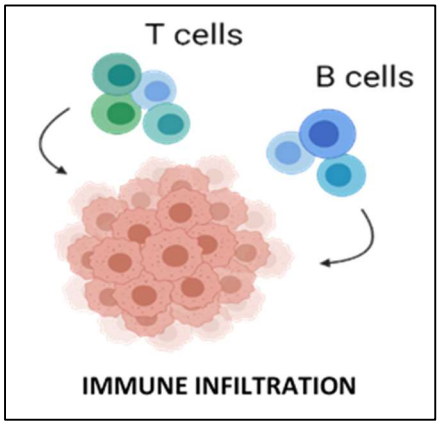

Figure 2. Immune infiltration of T and B lymphocytes in tumor microenvironment. Image created with Biorender.com

BCR (most commonly called as immunoglobulins, IG) consist of two identical heavy-chains (IGH) and two light-chains, Kappa (IGK) and Lambda (IGL). Human T-cell receptors (TCR) consist of an alpha and beta chains (TRA and TRB) and a gamma and delta chains (TRG and TRD). Antigen binding occurs in the variable domain, which is generated by recombination of a set of variable (V), diversity (D) and joining (J) gene segments forming the B- and T- cell immune repertoire (IR), and its diversity is mainly concentrated in the complementary-determining region 3 (CDR3)), from now on, this combination will be defined as V(D)J, (Figure 3).

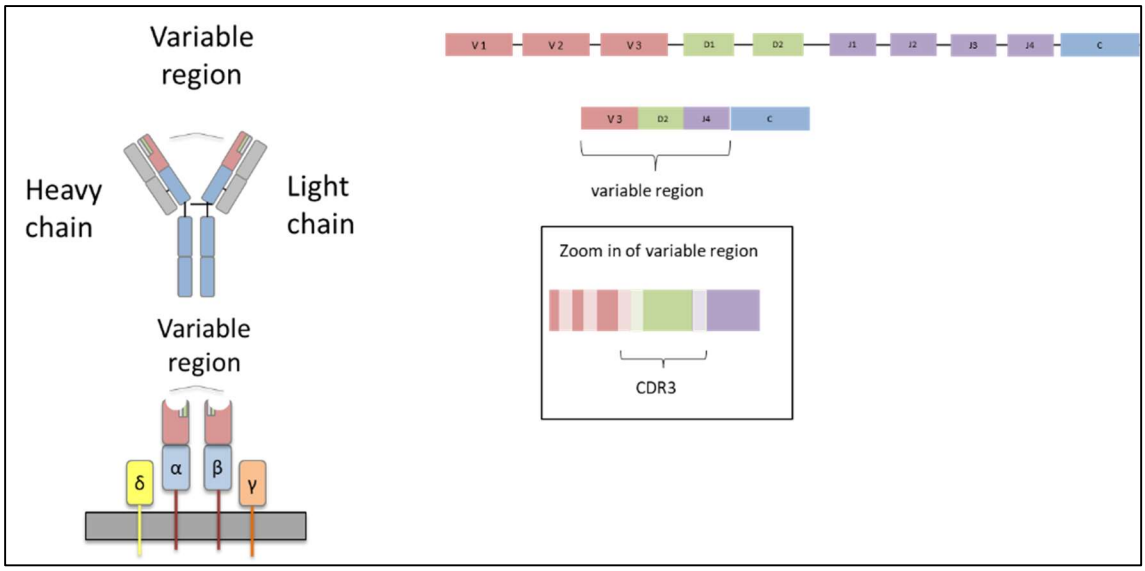

Figure 3. Diversity in antibodies and TCR are generated by the V(D)J recombination. Image created with Biorender.com

Tumor-infiltrating lymphocytes are associated with a favorable prognosis in breast cancer and non-small cell lung cancer and a bad prognosis with prostate cancer. However, the prognostic effect varies across grade and stage of tumors, clinicopathological factors and histological or molecular subtype[17].

In addition to their relationship with overall survival, immune infiltration should be taking into account along with other factors. On one hand, some of the known risk factors (diabetes, smoking, alcohol ABO blood group, and/or obesity) are involved in an inflammatory process which could be causing B and T cell infiltration in the tumor.

In a previous study, Pineda et al.[18] have analyzed IG/TCR richness and diversity finding that these measures are present in PC. Regarding its relationship with risk factors, it was observed a higher IG infiltration in heavy smokers and a larger TCR infiltration in individuals with previous history of diabetes. Additionally, they observed that individuals with high levels of IG infiltration had a better prognosis. However, there is still a high variability among individuals suggesting that other factors may play an important role in the explanation of this observed diversity. In fact, it is known that approximately 30 to 40% of the immune system differences are explained by genetic variants[19]. Therefore, we may consider that some of the variability observed in the intratumoral IG/TCR features could be explained by differences in the genetic susceptibility patterns across individuals.

1.3 Genetic Susceptibility

Since the Human Genome Project has emerged in 2003, many tools were available to manage large databases conformed by the reference human genome sequence and other relevant omics data. These tools have allowed researchers to improve their characterization of some diseases through the study of their association with determinate markers of genetic variations[20].

These selected markers are called single nucleotide polymorphisms (SNPs) which represents a change in only one nucleotide base pair position in the DNA chain. SNPs are the most abundant molecular markers in the genome and in consequence, they are chosen as markers for studying complex genetic traits and for understanding the genomic evolution[21].

Using SNP array technology is possible to cover a large variability of the genetic information from each individual. The majority of the SNPs are bi-allelic, indicating the two possible bases at the corresponding position within a gene. If we define A as the common allele and B as the variant allele, three combinations are possible: AA (the common homozygous), AB (the heterozygous) and BB (the variant homozygous). These combinations are known as the genotypes and they are assessed with SNP genotyping platforms [22].

Despite the wide range of benefits associated with genetic studies (gene markers discovery, medical advancements, genetic origin variation and rare genetic variants detection from a GWAS (Genome Wide Association Study) approach, among others) some limitations such as the low amount of variation of the phenotypes explained by the SNP, the difficulties in their analysis, and the high probability rate of false associations due to population stratification are important to consider. Some of these caveats could be solved using large sample size and the use of appropriate methodology[23].

Population Stratification (PS) is a common problem in the majority of the genetic studies due to the differences in allele frequencies between subpopulation. This problem could increase when there are two or more genetically distinct groups in a population. A clear example is when different races are included in the study and spurious correlations arise. In a typical case-control study with some people from European ancestry and other from Asian ancestry, a significant SNP could be associated with the disease of interest, but

indeed may be a spurious association since some specific SNP are just more common in the European population. Several methods are proposed to infer PS, including multidimensional scaling (MDS) and principal components analysis (PCA) which, based on linear models, represent visual methods to cluster samples by race[24]. The first five principal components are often used to correct PS bias although it is unclear if this procedure will actually mitigate that effect.

Along with PS effect, linkage disequilibrium[25] (LD), which refers to the association of alleles at different loci that occurs not at random, is employed as a statistical measure that compares haplotype frequencies (observed and expected) testing independence. Thus, it is common to filter those SNPs whose LD levels are higher than a selected cut-off point. This previous step is usually applied as a first variable selection process in order to avoid the multicollinearity problem presented in genetic analysis.

1.4 Methodological challenges using genetic data

The ultimate goal of this research is to investigate the role of the tumor immune infiltration in PC towards the identification of new risk factors, treatment options and prognosis improvement. To that end, integrative strategies are needed to combine molecular data very well characterize in datasets such as The Cancer Genome Atlas (TCGA)[26] with very well-defined epidemiologically and clinically phenotyped studies, such as the PanGenEU[27]. In this master thesis, we are planning to find the best model to predict the immunological infiltration using the SNPs as independent variables. With the signature obtained her, the immunological infiltration will be predicted in the PanGenEU to allow the identification of new exposure factors affecting the development of PC.

Genomic studies are complex and have several biases as previously commented, therefore the classical statistical assumptions are limited. In order to implement the models required for the analysis of this data, more advanced statistical techniques will be described and used in this work.

The most classical statistical approach to assess the relationship between genetic variants-SNPs (independent variables) with the IG/TCR measuring tumor immune-infiltration (dependent variables) is a linear regression model but, one of the main assumptions in a linear regression model is the independence between the regressors variables. In the present study, SNPs are considered as the independent variables, but as previously

explained, these variables might be highly correlated, especially if they are in the same gene closely located in the genome. Moreover, the high dimensionality is also a current problem affecting method convergence and being computational time-consuming.

To deal with such a number of variables, machine learning (ML) algorithms are proposed. These particular methods are based on two main phases: a training phase in which the algorithm learns from the given data and a predicting phase in which estimated values are calculated based on the learning process. Literature presents a wide range of machine learning methods showing a large variety of mathematical foundations to this first phase but, as for any study, the best approach will depend on the main study objective.

For example, regularized regression methods (Lasso, Elastic Net and Ridge Regression) are employed when there are many features (more than participants in the study, i.e., $p \gg n$) and a multicollinearity problem among them. Thus, some coefficients are sent towards zero reducing the variance of the model and doing a feature selection when that occurs. However, high throughput data have an extremely large number of variables and, sometimes, the selection of the ones that increase the predictive ability to the model is not straightforward. Therefore, the variable selection process, in machine learning algorithms is still an open question.

Approaches such as Random Forest and Neural Networks have been increased in popularity in the last few years. These techniques present algorithms (within ML) to correct overfitting and multicollinearity problem, too.

2. HYPOTHESIS AND OBJECTIVES

The hypothesis of this study is that there is already developed statistical methods to analyze the pertinence of using genetic susceptibility patterns to predict the variability of B and T cell repertoire variation of PC.

The main objective of this project is to identify which is the best statistical approach to deal with the multicollinearity problem and high-dimensional and complex genomic data. In particular, to predict richness and diversity of B and T cells receptors of PC using SNPs as independent variables. Thereby, four different machine learning methods in three different scenarios were benchmarked and compared in terms of prediction.

The specific objectives are:

I. To find different genetic susceptibility patterns associated with the tumor immune infiltration in PC to integrate in a future step this new characterization with non-genetic risk factors (tobacco, diabetes, asthma or allergies, among others).

II. To apply penalized regression methods in order to improve prediction and analyze different feature selection approaches.

III. To explore the differences between four machine learning approaches in this context.

3. MATERIAL AND METHODS

3.1 Data

In this project, a public database The Cancer Genome Atlas[26] (TCGA) has been used. This database consists of several omics data types (genome, transcriptome, methylome, etc.) measured in tumor tissue and genetic data measured in blood from 33 tumor types, including PC. In this project, 120 blood samples of PC patients were considered from which the RNAseq and SNPs genotyping data were available.

3.2 Variables

Previously to this work, the B-cell receptors or Immunoglobulins (IG) and T-cell receptors (TCR) were extracted using a well-known bioinformatic software MiXCR[28]. This tool aligns the raw RNAseq FASTQ files to the V(D)J recombination region (Figure 3) to extract IGH, IGK, IGL, TRA and TRB.

Richness and Diversity were calculated through Expression and Entropy measurements. Expression was estimated with the following formula:

$$Ig_i / TR_i = \frac{M_i}{N_i + M_i};\ i = 1,...,n \tag{1}$$

Where M_i is the number of reads that map to a specific VDJ recombination and N_i is the number of reads that map to a anything else in the genome in n samples.

And Shannon entropy (H index) was estimated as:

$$H = -\sum_{i=1}^{N} p_i \log_2(p_i);\ i = 1,...,N \tag{2}$$

N is the number of unique clones and p_i is the frequency of clone i.

Hence, the original dataset was formed by 120 samples (PC patients) with the values of ten dependent variables (Expression and Entropy of IGH, IGK, IGL, TRA and TRB) and demographical variables such as sex, age and race. Regarding to genotype data, there was a total of 906,600 SNPs which represents a huge number in comparison given small sample size.

Thus, our predictive model was built based on the set of Expression and Entropy as response variables and the selected SNPs as independent variables (inclusion criteria are described in Section 3.3):

$$Y \sim X + age + sex + error \tag{3}$$

Where Y represents each of the dependent variables (Expression and Entropy of IG and TCR) considered in this study and X the matrix with the independent variables (SNPs) adjusted by age and sex. Nevertheless, biological features such as SNPs variability and correlations among SNPs from the same chromosome along with the huge number of

variables were some issues to face up and because of this reason, several filtering steps were proposed (Figure 4).

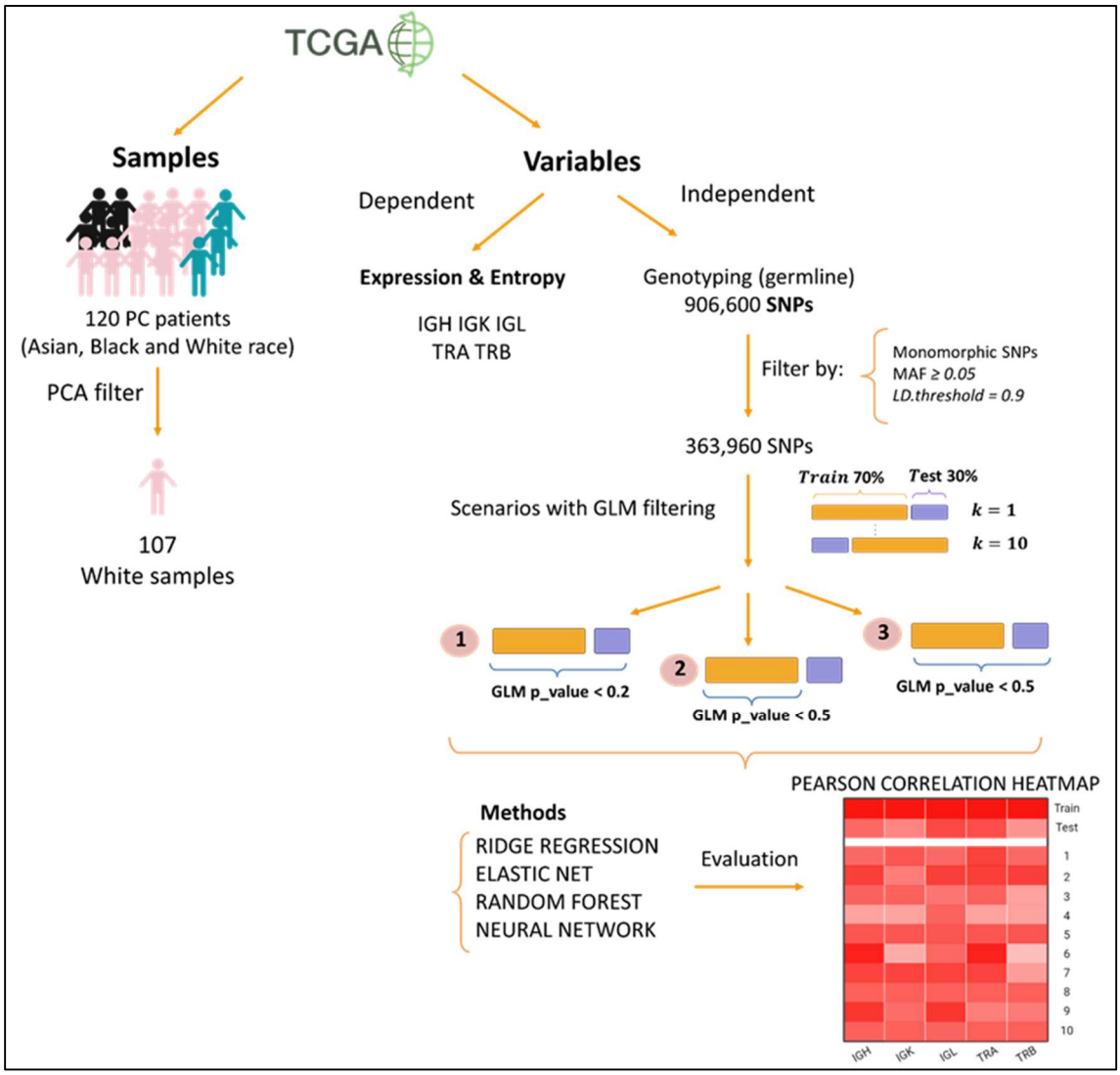

Figure 4. Flow diagram summarizing methodology with all filtering functions used in this work. PCA= Principal Component Analysis; MAF = Minor Allele Frequency; LD = Linkage Disequilibrium; GLM = Generalized Linear Model.

3.3 Filtering steps

- **Excluding monomorphic SNPs**

  Exploring genotype data, it was observed that there were some SNPs whose value did not vary across the samples. These SNPs, called monomorphic, did not give any information to the models. In fact, their variance was zero and in consequence, they were considered irrelevant variables. This is the reason why they were removed for further analysis.

- **Principal Component Analysis**

A frequent practice in genetic studies to show population stratification (PS) is through the plot resulting from a Principal Component Analysis[29] (PCA). In the dataset considered, there were individuals from three different races (Asian, Black or African American and Caucasian). Therefore, PS represented in the PCA (projection of the first two principal components and clustering patients by race) introduced a second filtering function, this time on samples, selecting only Caucasian individuals to prevent this systematic effect. In addition, some of the missing values could be classified in their corresponding category using this methodology.

- **Minor Allele Frequency & Linkage Disequilibrium**

Minor Allele Frequency (MAF) is defined as the frequency of the second most common allele in a given population. In this analysis, we excluded all the SNPs with a MAF < 0.05 since they were considered rare variants.

Calculation of this relative frequency is summarized in the following steps:

1) SNPs can take one of these three possible values: 0, 1 or 2 in each sample depending on the allele's combination observed in the individual (AA, AB and BB, respectively). Thus, three new variables were generated denoting the counts of these categories for each SNP.

2) Allelic frequencies were calculated considering the allele's load in each category:

$$p = \frac{2n_0 + n_1}{2n},\ q = 1 - p \tag{4}$$

where n_0 denotes AA allele combination, n_1 denotes AB allele combination, n_2 denotes BB combination and $n = n_0 + n_1 + n_2$ number of total samples participated in the study. Therefore, p is referred to A allele frequency and q to B allele frequency.

3) Last step consisted in determining which frequency referred to the minor allele frequency, that was the minimum value between p and q.

LD avoids multicollinearity and convergence problems when a determined threshold is established. Briefly, LD calculation is based on r^2 coefficient which was obtained as follows:

$$r^2(p,q,p_{AB}) = \frac{(p_{AB} - pq)^2}{p(1-p)\times q(1-q)} \quad (5)$$

being p_{AB} frequency of *AB* haplotype and *p* and *q* defined as in (4). Therefore, when high r^2 (>0.9) between two SNPs is observed, one of them is randomly excluded to prevent an excessive correlation between SNPs that were closely related in the genome.

- **Generalized Linear Model**

  High dimensionality is a current challenge observed in our data. For that reason, many studies applied a filtering step to reduce the number of SNPs to be introduced in the prediction models. Usually a univariate generalized linear model (GLM) for each SNP is fitted. Due to the lack of a gold standard cut-off to determine which SNPs are filtered, different scenarios were proposed in this study to find the one with the better prediction accuracy.

3.4 Methods

Given the nature of the variables under study, it is important to apply the appropriate methodology that better fits to our high dimensional and complex genomic data while dealing with the multicollinearity problem, a key issue in this project. Thus, a total of four methods were considered: Ridge Regression (RiR)[30] and Elastic Net (ENET)[31] Random Forest (RF)[32–36] and Neural Network (NN)[37–39]. By this way we were able to tackle the problem from different approaches elucidating which kind of methodology provided the most predictive ability in our data.

3.4.1 Ridge Regression & Elastic Net

Unlike traditional regression, penalized regression methods can usually deal with all the challenges associated with high throughput omics data so that they are able to fix the $p \gg n$ challenge through different approaches. Then, within regularized regression, we distinguish three main methods:

- Ridge Regression proposed by Hoerl and Kennard[30] solves collinearity problems by including a penalty term that contracts the regression coefficients. However, it is not suitable when there is a high number of p variables, since no variable selection occurs.
- Lasso Regression proposed by Tibshirani[40] allows a variable selection, unlike RiR, by giving some coefficients an estimation of a 0 value. A valuable advantage of these models is that since they are parsimonious, their interpretation is less complex.
- Elastic Net Regression results from the combination of Lasso Regression and RiR, taking advantage of the benefits offered by both methods and overcoming some of their drawbacks. This new technique proposed by Zou et al.[31] is able to solve correlation problem by choosing groups of variables (nets) and to work with a larger number of variables than observations ($p \gg n$).

As defined in Section 3.2, our matrix was made up of almost one million SNPs and 107 samples, therefore, we were in the $p \gg n$ scenario. Two different approaches, RiR and ENET, were applied and compared to address the main objective. Thus, the feature selection performed by ENET will be assessed in this particular biological context.

From this point, a general definition was described referring to both methods and identifying the main differences between these two methods.

Mathematical definition

ENET performs a variable selection based on the standardization of the p predictors (mean 0, variance 1) and a centered response:

$$\sum_{i=1}^{n} y_i = 0,\ \sum_{i=1}^{n} x_{ij} = 0 \text{ and } \sum_{i=1}^{n} x_{ij}^2 = 1,\ \text{for } j = 1, 2, ..., p \tag{6}$$

where $y_i = (y_1, ..., y_n)$ is the response vector in n samples and its predictors variables associated are denoted by $x_{ij} = (x_{i1}, ..., x_{ip})^T$. Thus, we can define $\hat{\beta}$ estimator for Elastic Net as follows:

$$\arg\min_{\beta} \{L(\lambda_1, \lambda_2, \beta)\} = \| y - X\beta \|^2 + \lambda_2 \| \beta \|_2^2 + \lambda_1 \| \beta \|_1 \tag{7}$$

where $|| \beta ||_2^2 = \sum_{j=1}^{p} \beta_j^2$ and $|| \beta ||_1 = \sum_{j=1}^{p} | \beta_j |$, λ_1 and λ_2 are hyperparameters that control penalization grade so that when the larger they are, the less value are taken by the predictors (major penalty).

Hence, coefficients vector is normalized with L_1 and L_2 norms, and shrinkage some of the coefficients towards 0.

If we define $\alpha = \lambda_2 / (\lambda_1 + \lambda_2)$ we could rewrite the formula identifying an optimization problem where the objective function is:

$$\begin{aligned} &\hat{\beta} \times \arg\min_{\beta} || y - X\beta ||^2 \\ &\text{subject to: } \alpha ||\beta||_2^2 + (1-\alpha) || \beta ||_1 \leq t \text{ for some } t \end{aligned} \tag{8}$$

In this way, RiR is performed when $\alpha = 1$ and Lasso when $\alpha = 0$. Thus, ENET represents a combination of both penalized regression methods:

$$\begin{aligned} \arg.\min_{\beta_{Ridge}} \{ L(\lambda_2, \beta) \} &= || y - X\beta ||^2 + \lambda_2 || \beta ||_2^2 \\ \arg.\min_{\beta_{Lasso}} \{ L(\lambda_1, \beta) \} &= || y - X\beta ||^2 + \lambda_1 || \beta ||_1 \end{aligned} \tag{9}$$

RiR, unlike ENET and Lasso, is not a variable selection method, since it does not exclude any variables from the model.

Procedure

To summarize RiR and ENET penalized regression algorithms, we can establish the following steps:

1) An alpha vector was defined taking the unique value of 0 when RiR was applied and a sequency from 0.1 to 0.9 when ENET was considered.

2) For each alpha, a *k*-fold cross validation was performed and lambda values were calculated in each iteration. Then, mean square error (MSE) was estimated between observed and predicted values obtained with such lambda.

3) The lambda obtained the minimum MSE was chosen to predict in a testing dataset in the case of RiR (unique alpha). In ENET, as several alphas were used in model fitting, the alpha with the least MSE was selected along with its lambda to predict in the testing set.

3.4.2 Random Forest

Introduced by Breiman[32], RF is defined as an ensemble of trees generated with random vectors identical and independently distributed that differ according to the chosen approach in the tree building process. In particular, bagging, random split selection and random subspace method are three of the main approaches used in this field.

Mathematical definition

This algorithm builds up multiple independent decision trees and combines them in order to obtain the best and robust prediction that reduces its variance:

First, for each tree, a bootstrap sample is generated from the training data. Next, a decision tree is growth to the bootstrapped sample with a feature selection of these variables and then, when iterations end, the random forest is given as output with the predicted values of the samples.

In classification, the class with the greatest number of votes is selected as predicted value and in the case of regression, the predicted values are an average of all the predictions in each regression tree.

$$\text{Regression} \rightarrow \hat{y} = \frac{1}{B}\sum_{b=1}^{B} T_b \text{ for } b = 1, ..., B \tag{10}$$

$$\text{Classification} \rightarrow \hat{y} = \text{majority vote}\left\{\hat{C}_b\right\}, \text{ for } b = 1,...,B \tag{11}$$

Where $\hat{y}$ represents the predictions in each case, B is the number of random forests, T_b and $\hat{C}_b$ are respectively the random tree and the class prediction in the *bth* random-forest tree.

Out of Bag Error

RF algorithm is based on the principle of grow as many decision trees as subsamples have been generated. In such a way that, in each of them the model is built by a specified number of individuals of the training set and assessed on the remaining subset. Samples that take part in this evaluating set, are called out of bag samples and associated to them, it is estimated an error known as out of bag error.

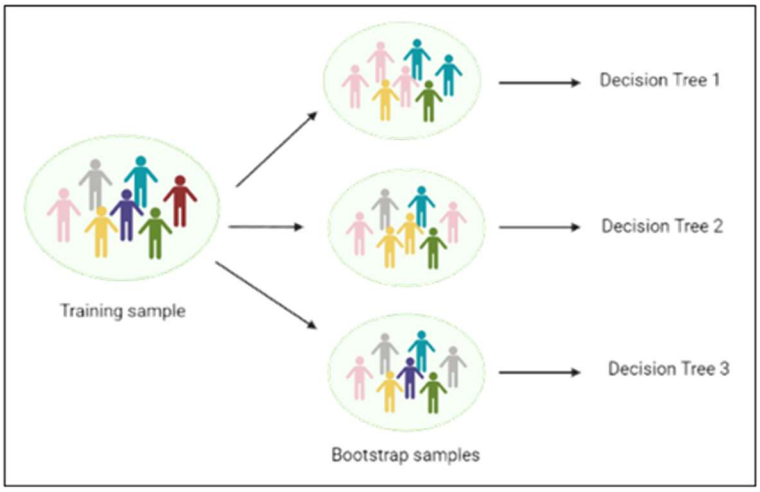

Figure 5. Random Forest model with bootstrap samples. Image created with Biorender.com

Hyperparameters

Machine learning algorithms are characterized by a range of values configuration which have been defined previously by the user. They are called hyperparameters and usually there is not a gold standard criterion to select the proper ones. Therefore, it is necessary to make some iterative tests in order to obtain the optimum value. The most important hyperparameters in RF are:

– M: number of predictors chosen as candidates in each partition. Hence, being p the total number of predictors, $M < p$. Depending on the problem to solve, M often takes the value of $\sqrt{p}$ for categorical response variables and $p/3$ for continuous response variables. However, better predictors can be obtained with other values than with default ones.

– B: number of trees to build RF. Thus, the bigger B is, the higher computational cost will be. Additionally, the number of trees should be determined according to the number of samples (n) in order to avoid an overfitted model.

3.4.3 Neural Network

Another machine learning technique, in the context of supervised algorithms, is NN. The way it works is similar to neurons of nervous system so that a set of neurons (or nodes) forms the input layer and are connected with other neurons belonging to a hidden layer throughout some weights associated with the first inputs. By this way, like the human brain, information is transmitted from one neuron to another until the last output layer is reached.

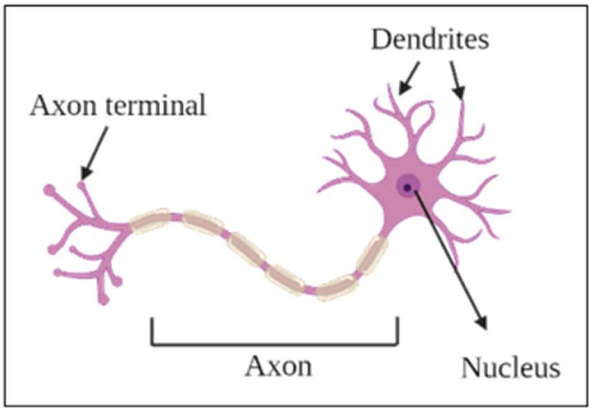

Figure 6. Biological neuron. Image created with Biorender.com.

Mathematical definition

We can define this information transmission with the equation as follows:

$$h_i = f\left(\sum_{i=1}^{n} w_i(t)x_i(t) + u_i\right) \tag{12}$$

where h_i is the output layer, $f()$ is the activation function, w_i are the weights of that layer associated to x_i neuron and u_i is the bias that can be simplified initializing the function in $i = 0$ so as to determine $w_0(t)x_0(t)$ where $x_i(t) = 1$.

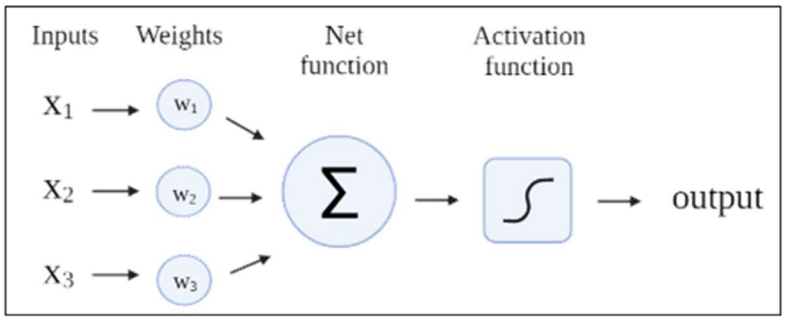

Figure 7. Process element diagram. Image created with Biorender.com.

Architecture

Related to its configuration, the most popular and widely used model is multilayer perceptron in which the network is configured by an input layer, an output layer and a one or more hidden layers.

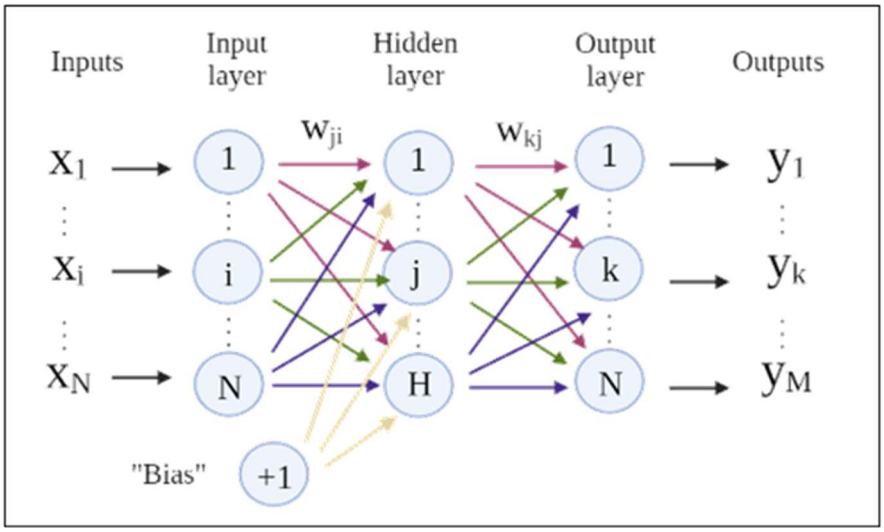

Figure 8. Multilayer perceptron (N-H-M) with one hidden layer, N input neurons, H in the hidden layer and M outputs. Image created with Biorender.com.

In this particular model, there is always a forward information transmission, that is to say from the input layer to the output layer. The N number of neurons in the input layer is determined by the number of predictor variables and, the number of neurons in the output layer will take the value of 1 if we are in the case of a continuous response variable (as it is presented in this study) or, in the case of a classification problem, each neuron will represent a category with a maximum and minimum values that delimit the predicted category.

Activation function

It has defined previously that in order to obtain the hidden layer output, it is necessary to apply an activation function. This is an infinite domain bounded function applied to weighted and summed inputs to limit the amplitude of the output signal. Frequently, it is an increasing monotonic and continuous function and, despite the fact that there are several types of it, sigmoidal and ReLU functions are most generally used in regression and classification problems:

- Sigmoidal activation function (classification problems): increasing function with asymptotic properties.

$$y = \frac{1}{1+e^{-x}} \tag{13}$$

being y and x the dependent and independent variables respectively.

- Rectified Linear Unit (ReLU) function (numeric prediction problems): function with a range from 0 to infinity.

$$y = \max(0, x) \tag{14}$$

Activation functions have the role to decide which neuron is activated according to its information relevance. However, depending on the problem to solve it is more appropriate to use one function or another. Sigmoidal activation is the most appropriate function in classification problems since its values range from 0 to 1 and its output it is considered as a probability. On the other side, ReLU activation function fits better to regression problems when these values have the same range as the function (no negative values).

Internal validation in Neural Network

In the previous sections, it was mentioned the existence of some weights associated with each of the neuron components of the net, however, their initial value and their actualization correspondent to each layer have not been determined yet.

It is in the training phase when the weights changed their values by backpropagation techniques to minimize the loss function:

$$E = \frac{1}{S}\sum_{s=1}^{S} E_s \text{ for } s = 1,...,S \tag{15}$$

where S is the number of training set repetitions and E_s the associated error to the *sth* repetition.

Mathematical foundation of backpropagation algorithm is based on gradient descent method where weights are updating according to a learning rate η and the information about the direction that partial derivative provided. Hence, for the w weight, its update is given by the following expression:

$$\Delta w(i) = -\eta \frac{\partial E}{\partial w}[s] \tag{16}$$

In this training phase, it should be noted two relevant concepts: epochs and batch. Epochs are defined as the number of times the dataset is introduced in the NN. The more diversity is observed in our data, the larger the number of epochs needed but, may also lead to an overtraining problem. Related to batch term, training set is divided in several parts or batches, thus, an epoch will have as many iterations as batches are observed in the training set. The repetition set that makes up each batch, enters to the neural net before weight updating achieving a convergence acceleration.

Once these concepts have been defined, we can summarize the training phase in four steps:

1) Initialize the weights from a random mode or through normal or uniform functions.
2) Split the data in a number of batches and make a first estimation for the first epoch.
3) Calculate the loss function E that measures the differences between predicted and observed values within the training dataset.
4) Minimize the loss function using backpropagation techniques with the gradient descent algorithm.

In conclusion, weights are updated depending on the minimization of E each time the set of data is passed to NN.

3.5 Training and testing processes

To evaluate methods performance, the whole dataset was divided into two distinct subsets:

- A training set (70% of the total sample) in which all methods fitted their functions and determined the best predictors to estimate response variables.
- A testing set (30% of the total sample) where the same predictors used in the learning process are employed to estimate our variables of interest.

Once predictions were estimated, as they are continuous and numeric, Pearson correlations (ρ) were calculated in both training a testing set. Therefore, observed and prediction values were compared measuring its relationship grade with this coefficient.

3.6 Software

All methods and previous pre-processing filters applied in this project were executed with statistical software R[41] version 4.0.2 and, with it, graphics presented in results section were done with this software. Then, for each method a distinct package was installed and the correspondent function was run (Table 1):

***Table* 1.** Packages and functions applied in methods

	ENET & RiR	**RF**	**NN**
Package	*Glmnet*	*randomForest*	*keras*
Function	*cv.glmnet*	*randomForest*	*fit*

Machine learning algorithms are described by their multiple hyperparameters. Thus, each function is defined explaining its main arguments and deciphering the reason why the established value was chosen.

3.6.1 Elastic Net & Ridge Regression

Rcode 1. ENET & RiR function

```
cv.glmnet(train, y_train, family = "gaussian", alpha = alpha,
nfold=3, parallel=F, standardize=TRUE, type.measur='mse')
```

ENET and RiR were applied with the same function since RiR represents a particular case of ENET where alpha parameter is equal to 0. This function introduces independent and dependent variables of the training set (*train*, *y_train*) as argument, identifying a gaussian family because of the numerical nature of dependent variables. Also, another parameter to comment is *nfold* which is the number of folds selected for the CV procedure to find the optimal lambda parameter. This number was established as the minimum (*nfold* = 3) in order to avoid an overfitting problem since the training sample size was quite small. Additionally, a standardization of the variables was performed and MSE was chosen as the main measure to compare the best value of alpha vector in ENET.

Finally, predicted values were obtained with *predict* function which returns predicted values using the same variables as the ones chosen by *cv.glmnet* objet (*fit*) and employing the minimum lambda selected in the training phase which represents the value with the minimum MSE.

Rcode 2. Prediction Train and Test functions in ENET & RiR

```
pred_train = predict(fit, newx = as.matrix(train),
                     s = "lambda.min")
pred_test = predict(fit, newx = as.matrix(test),
                     s = "lambda.min")
```

3.6.2 Random Forest

Rcode 3. RF function

```
randomForest(x= TrainSet[,-1], y=TrainSet[,1], ntree = 500)
```

As in ENET and RiR, the training set was passed to the function (where *x* represents regressors variables and *y*, the outcome variable of interest). In RF, an important parameter to determine is the number of trees (*ntree*) which was chosen as 500 based on the large sample size and considering to produce a stable model. Furthermore, *mtry* hyperparameter which does not appear in the above function syntaxis is settled by default with *p/3* (*p* = number of predictors).

After model was fitted in the training phase, predicted values were calculated, both in training and testing sets:

Rcode 4. Prediction functions in RF

```
predict(model[[i]],TrainSet)
predict(model[[i]], ValidSet)
```

3.6.3 Neural Network

To introduce NN function, three main steps explain deep architecture, compilation and fit:

1) Architecture:

Rcode 5. Architecture definition in NN function

```
model <- keras_model_sequential()

model %>%
layer_dense(units = 512, activation = 'relu', input_shape =
c(ncol(x_train))) %>%
layer_dense(units = 256, activation = 'relu') %>%
layer_dense(units = 128, activation = 'relu') %>%
layer_dense(units = 1)
```

A total of 5 layers determined NN structure starting with an input layer with as many neurons as SNPs analyzed in the data (*input_shape*) and reducing this large number of neurons the three next hidden layers, being the output layer one unique neuron. Moreover, a rectified linear unit activation function was present in all layers (*activation* = *relu*).

2) Compilation:

In addition to the activation function and the number of neurons participate in the net, it was needed a learning rate (*lr*) which controls that not many oscillations result in the process. Mean absolute error was chosen in this method to evaluate the training process across 20 epochs defined in the fitted model.

Rcode 6. Compilation function in NN

```
model %>% compile(loss = 'mae',optimizer=optimizer_adam(lr =
0.00001),metrics =list("mean_absolute_error"))
```

3) Fit:

For the fitted model, in a sample size of 75 patients (number of rows in the training set), dataset was split into 7 batches, each with 10 samples (*batch_size*). In one epoch, 7 baches were performed and, as there were 20 epochs, a total of 140 batches were passed during the entire training process.

Rcode 7. Fitted function in NN

```
model %>% fit(x_train, y_train, epochs = 20, batch_size = 10)
```

Then, like with the two other methods explained before, predicted values in the training and testing sets were estimated with the fitted model:

Rcode 8. Prediction functions in NN

```
Y_predTrain[[i]] <-model %>% predict(x_train)
Y_predTest[[i]] <-model %>% predict(x_test)
```

3.7 Scenarios

After applying monomorphic, MAF and LD filters, the number of SNPs was reduced from a total of 906,600 variables to 363,960. Although it represented a considerable reduction step in our complex genomic-data, there was still observed a high dimensionality which could prevent machine learning methodologies to obtain a favorable performance especially given the limited sample size (107 Caucasian patients). Thus, a GLM was proposed to diminish the large number of SNPs.

Relating this GLM filtering function, three scenarios were contemplated due to the influence that the SNPs selected have in the prediction results. In all scenarios, the dataset was split in a training set of 70% of the whole samples and 30% for the testing set (this is 76 individuals participate in the training and 31 in the testing) repeating this process randomly 10 times. Nonetheless, depending on the GLM approach this partition was carried out before or after the selection SNPs step.

Scenarios proposed in this project are the following:

1) The whole dataset was employed to perform a univariate GLM selecting those significant SNPs with a p-value < 0.2. Then, the dataset was split randomly in training and testing across 10 iterations and, methods accuracy was benchmarking in the remaining testing set.

2) Samples were separated in 10 different training and testing sets and GLM was performed only in the training sets. Therefore, ten models were estimated (one for each iteration) and diverse SNPs were chosen according to p-value$<$ 0.5 to take

part in the machine learning methods. Finally, as in the first scenario, the prediction accuracy was compared in the testing dataset.

3) The last scenario considered was a combination of the other two above-mentioned cases where the GLM was applied using all the samples but with a less restrictive cut-off than the one chosen in the first scenario, the same used in the second one (p-value$<$ 0.5). This means that methods are doing the predictive analysis with the significant SNPs plus a considerable amount of noise (SNPs that were not associated with the response variables).

4. RESULTS

4.1 Descriptive Analysis

Population descriptive analysis is presented in the following tables where sample distribution of categorical variables is shown in Table 2 and summary statistics for continuous variables is shown in Table 3:

Table 2. Data distribution by sex and race

		Global N = 120	**Caucasian N = 103 (85.8%)**	**Rest N=13 (10.8%)**	**Missing N=4 (3.3%)**
Sex	Male	63 (52.5%)	57 (55.3%)	5 (38.5%)	1 (25.0%)
	Female	57 (47.5%)	46 (44.7%)	8 (61.5%)	3 (75.0%)

The whole dataset was characterized by a balanced sex population (with almost equal proportions for males and females in global dataset and in filtered Caucasian race) and a high proportion of Caucasian patients representing around 86% of the 120 individuals. The median of the age is 66 years old in both groups (Table 3).

Table 3. Summary of the continuous variables (age, IGH, IGK, IGL, TRA, TRB, Expression and Entropy)

		Global		**Caucasian**	
		Median	**IQR**	**Median**	**IQR**
Age		66	16	66	15
Expression	IGH	0.0002	0.0004	0.0001	0.0004
	IGK	0.0002	0.0004	0.0001	0.0004
	IGL	0.0001	0.0003	0.0001	0.0003
	TRA	0.0000	0.0000	0.0000	0.0000
	TRB	0.0000	0.0000	0.0000	0.0000
Entropy	IGH	7.4308	2.5250	7.3273	2.3542
	IGK	6.3196	0.8047	6.3139	0.6963
	IGL	6.3079	1.1742	6.1712	1.0924
	TRA	4.1279	2.2120	4.1683	2.1032
	TRB	3.9406	1.8458	3.9754	1.7833

IQR: Inter quartile range

Regarding the variables of interest, Expression values have very low values in comparison with Entropy, although a logarithmic transformation was considered in the next steps of the study. On the other hand, Entropy measures have different ranges depending on the studied variable. TCR (TRA and TRB) were characterized by a low median (around 4 in global and Caucasian individuals) but with a similar range as IG. Within IG Entropy, IGH was the variable with the highest median and the widest interquartile range (IQR = 2.525 in Global dataset and IQR = 2.3542 in Caucasian individuals).

4.2 Exploratory Analysis

As a first step, a widely known practice in genetic studies is to represent sample distribution depending on their race in order to detect population stratification. It can be studied through a Principal Component Analysis (PCA) and represented with the typical PCA plot including the two first principal components in the main axis. However, before performing this analysis, monomorphic SNPs (those which have the same value across all the samples) were removed in order to filter variables that did not give any information in the analysis (their variance would be 0 in the PCA). Then the total amount of SNPs was reduced to 883,231 which means an important reduction filtering function.

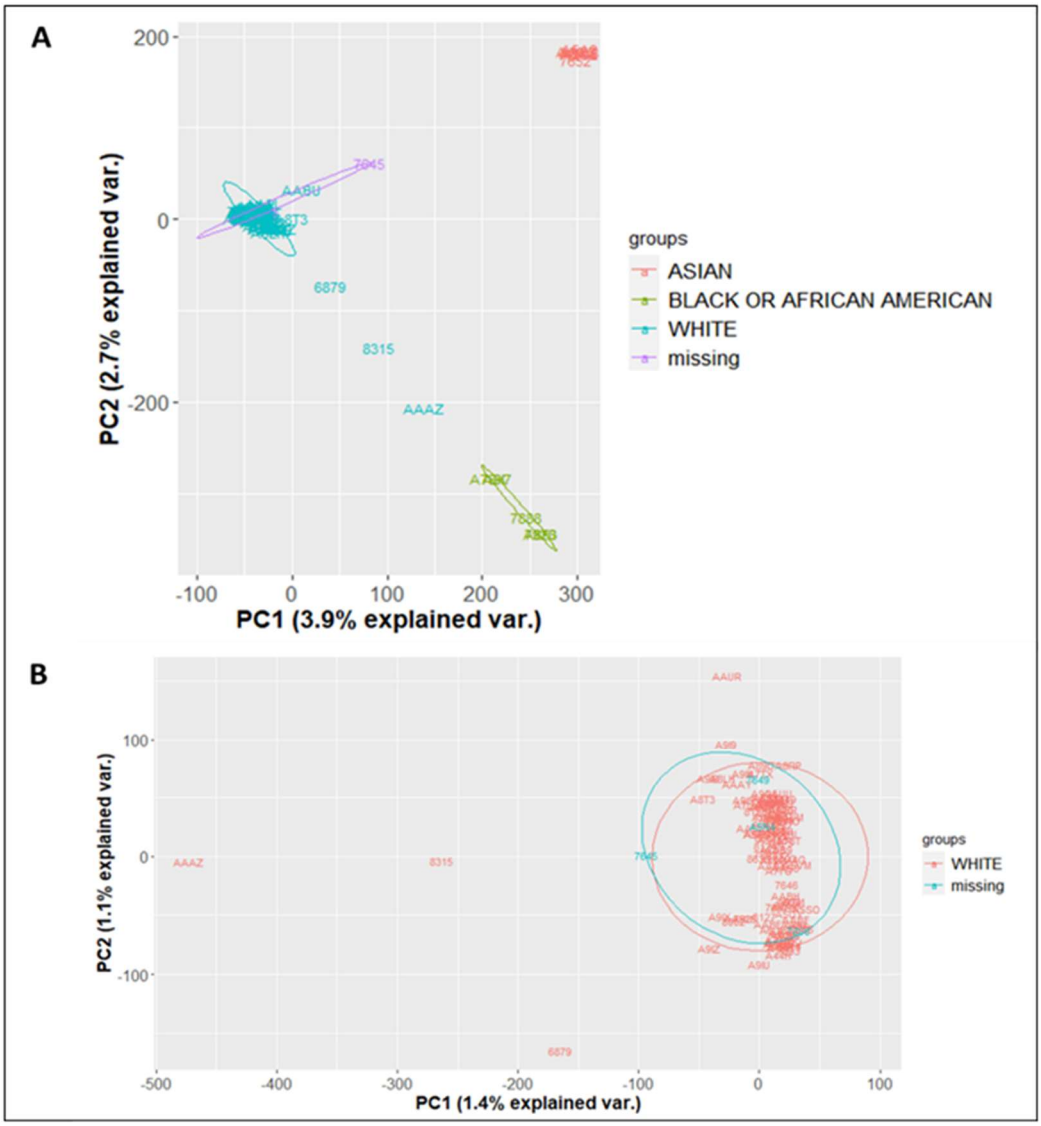

Figure 9. Principal Component Analysis (PCA) showing sample distribution by race in all individuals (A) and considering only Caucasian individuals and missing (B). PC1 and PC2 represent the first two components in PCA

In the PCA plot presented in Figure 9A, despite the small percentage of explained variance by the two first components, individuals were grouped by race in three different clusters. Furthermore, thanks to this representation, it was possible to categorize race missing data as Caucasian individuals since they are very close to these points. For further analysis, to avoid population stratification (PS) only Caucasian individuals were selected. Thus, 107 individuals took part in the following analysis.

In consequence, the same procedure was repeated with Caucasian and missing individuals (categorized henceforth as Caucasian population). Now, PCA plot (Figure 9B) presents how patients of this particular race were grouped together except from some outliers far from that circle area.

However, as observed in the PCA plot with all samples, the variance explained by the two first components was not large. In fact, there was a quite low proportion of variance explained for the 5 first principal components. Even if we considered all together, their sum did not reach 6% of total variability (Table 4).

Table 4. Summary of the first 5 components in the PCA including its standard deviation, proportion of variance and cumulative proportion

	PC1	**PC2**	**PC3**	**PC4**	**PC5**
Standard deviation	58.96	52.87	52.07	51.84	51.47
Proportion of Variance	0.0139	0.0112	0.0108	0.0107	0.0106
Cumulative Proportion	0.0139	0.0251	0.0359	0.0466	0.0572

As it was explained in Section 3.3, two other biological filters were considered to avoid methodological problems and consequently reduce dimensionality which was a current problem in this analysis. One of the filters regarded with the minor allele frequency (MAF) function which was applied to exclude rare SNPs that are presented in less than a 5% of the population sample (Table 5). The other filter regards to SNPs with a cut-off of 0.9 in linkage disequilibrium (LD). These SNPs were discarded since it was observed a correlation higher than 0.9 (the function retains randomly one of the SNPs in which it was observed that correlation).

Table 5. Minor Allele Frequency calculation in an example of 4 SNPs

	n	**n_0**	**n_1**	**n_2**	**p**	**MAF**
SNP_A-1780270	107	24	49	34	0.4533	0.4533
SNP_A-1780271	107	68	33	6	0.7897	0.2103
SNP_A-1780272	107	3	19	85	0.1168	0.1168
SNP_A-1780274	107	43	53	11	0.6495	0.3505

MAF = Minor Allele Frequency

Considering these previous filters, a univariate GLM was performed for each SNP adjusting by sex and age to find which SNPs were significantly associated with each variable of interest (Expression and Entropy of IGH, IGK, IGL, TRA and TRB).

One of the most important assumptions in a GLM is the normality distribution of the dependent variable. A good way to check normality hypothesis is density plot with Kolmogorov Smirnov (with Lilliefors correction) (K-S-L) test for normality (since

number of samples is greater than 50). In Figure 10, the density plots of the distribution of each dependent variable are represented and in Table 6, K-S-L test p-values are reported for each measure of interest.

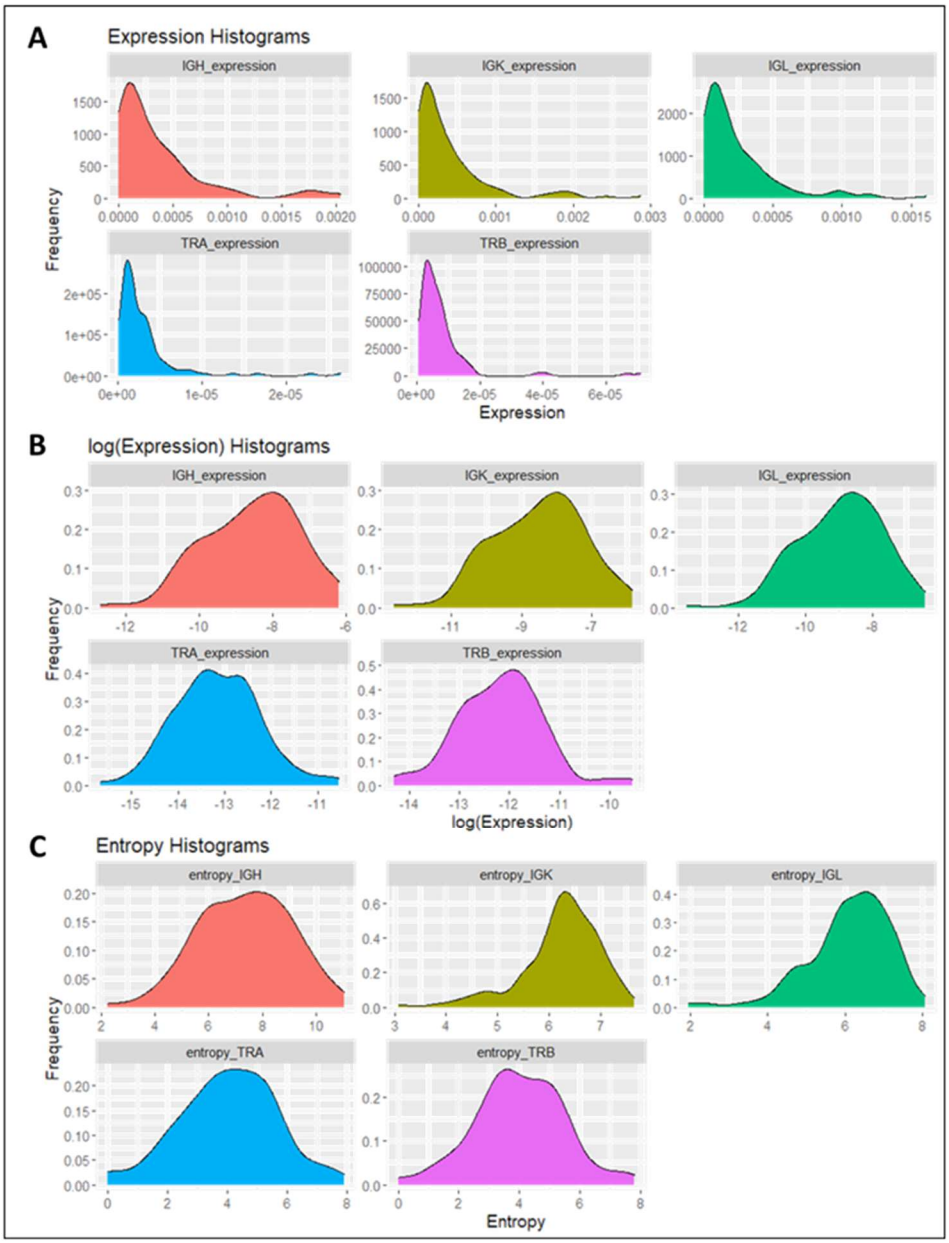

Figure 10. Density distributions for variables of interest (IGH, IGK, IGL, TRA, TRB) in each measure (log(Expression) and Entropy)

To accept normality assumption in Expression variables (Figure 10A) a logarithmic transformation was considered in order to standardize their values and to get a more homoscedastic value (Figure 10B). In Entropy variables (Figure 10C), no transformation was considered since they seem to follow a normal distribution.

Table 6. Kolmogorov Smirnov (Lilliefors correction) p-values test for variables of interest (IGH, IGK, IGL, TRA, TRB) in each measure (log(Expression) and Entropy)

	IGH	IGK	IGL	TRA	TRB
log(Expression)	0.3108	0.3721	0.2245	0.7888	0.7388
Entropy	0.5469	5.574e-05	0.01431	0.4334	0.7642

In log(Expression), all IG and TCR were accepted to follow a normal distribution (p-values > 0.05). However, for IGK and IGL Entropy, p-values were less than 0.05 which means that for both variables, normality assumption was rejected. Furthermore, selecting a random sample of 8 SNPs, we checked the normality assumption of these two variables based on the obtained residuals, (Figure 11 shows one randomly selected SNP and Appendix pages 1-2 eight selected SNPs for both variables). We could observe that except from the tails, both variables followed a normal distribution.

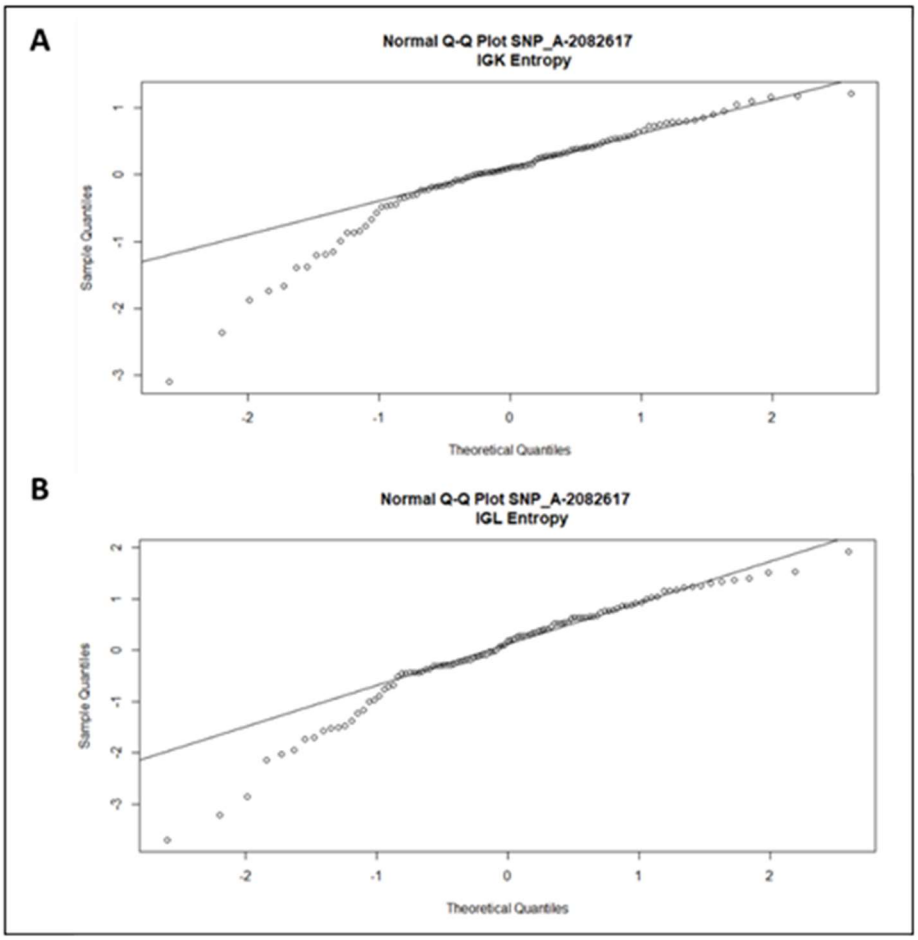

Figure 11. QQ plots of fitted model of one randomly selected SNP considering IGK Entropy (A) and IGL Entropy (B) as outcomes

Therefore, the same next steps were applied to all variables of interest in order to standardize the process.

Hence, a GLM univariate model (SNP by SNP) was built for each variable and in each measure to quantify how many SNPs were associated with these variables and the intersection among them:

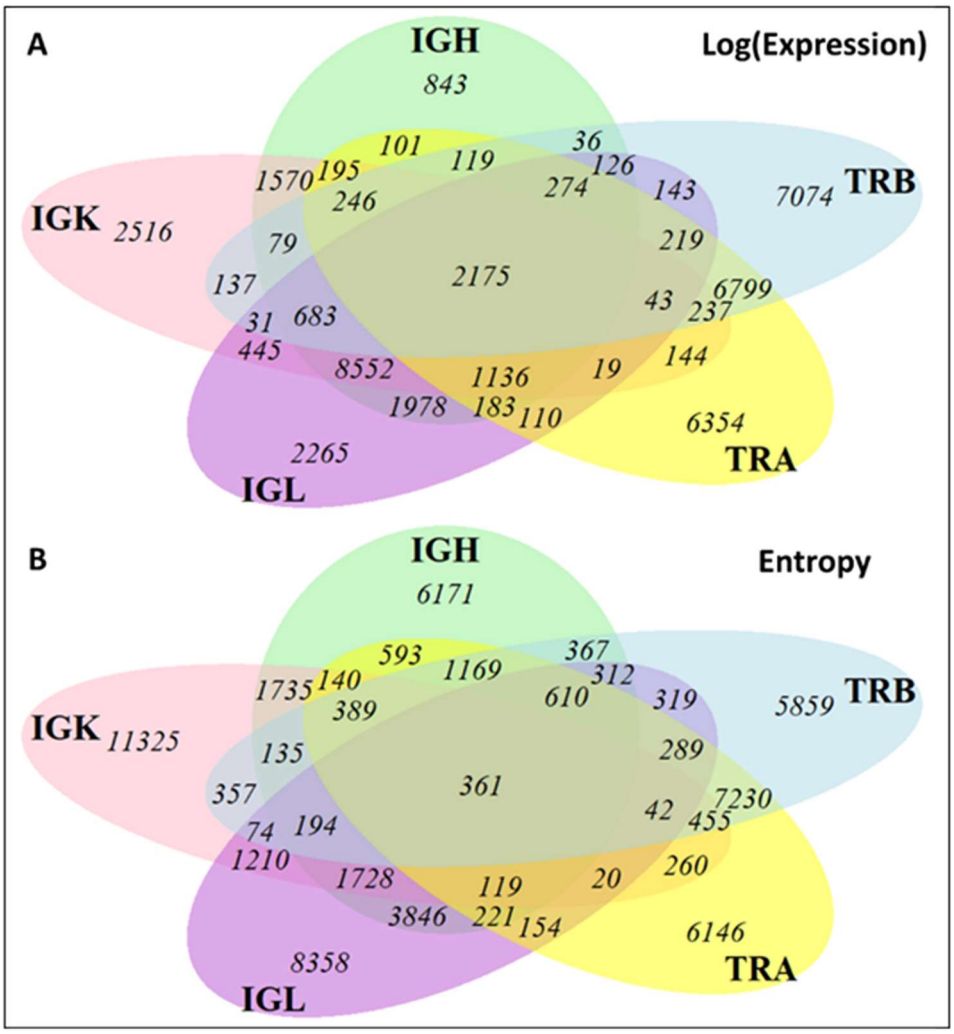

Figure 12. Venn diagrams showing significant SNPs in Expression (A) and Entropy variables (B) along with the intersection among B-cell receptors (IGH, IGK, IGL) and T-cell receptors (TRA and TRB)

GLM results were shown through Venn diagrams (Figure 12) where although some SNPs were overlapped between Expression and Entropy variables, there was a large number of significant SNPs (p-value < 0.05) not shared between them and being specific of the two types of receptors. The same occurs when both measures were compared. However, in this case, it was observed a bigger intersection in TCR than in IG (Figure 13):

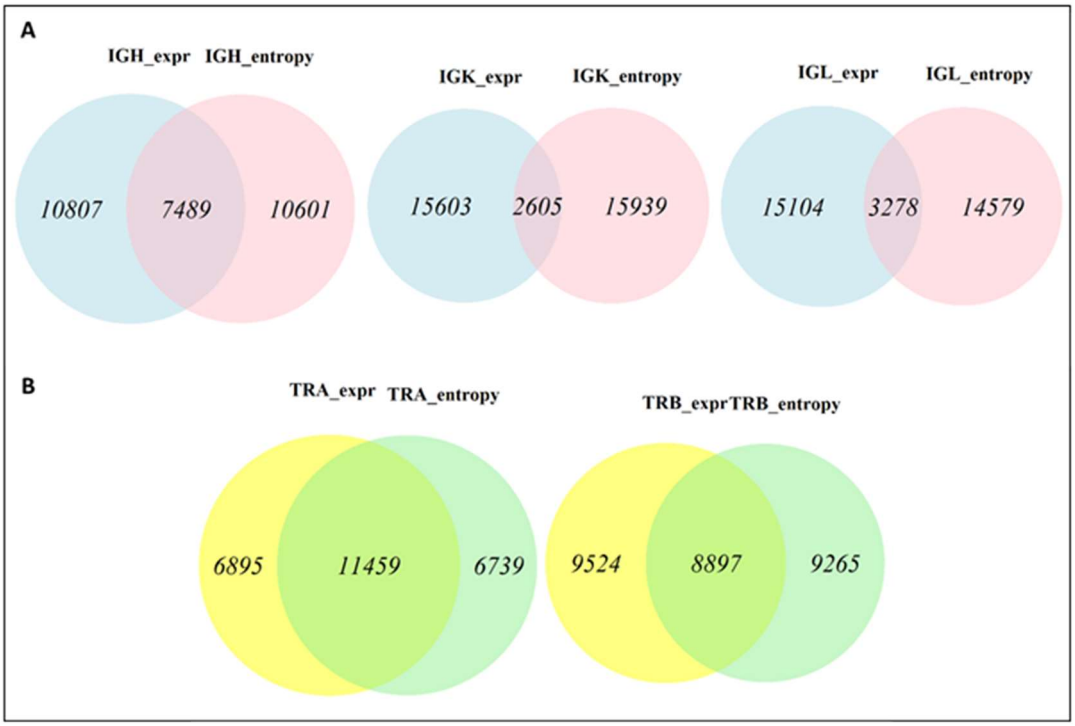

Figure 13. Venn diagrams representing overlapped SNPs between expression and entropy in IG (A) and TCR (B)

For example, when focusing on IGK (B-cell receptor) and TRA (T-cell receptor) it was clear that the intersection between SNPs associated with IGK is much smaller than the ones with TRA where the number of shared SNPs represented almost two times the unique SNPs observed in these measures. Statistically, this is a likely phenomenon because the correlation between IGK Expression and Entropy was lower than the TRA Expression and Entropy. Biologically, this could be because the IG was more diverse than TCR in pancreatic tumors.

4.3 Assessment of the predictive accuracy in each scenario

Results obtained in each scenario are shown below through heatmaps representing the absolute Pearson correlations between observed and predicted values across 10 iterations (rows in the heatmap) in the testing set for the 5 different variables of interest: IGH, IGK, IGL, TRA and TRB in log(Expression) and Entropy. The average of both, training and testing sets are represented in the two first rows.

4.3.1 Scenario 1: GLM p-value < 0.2 in the whole dataset

In Table 7 the number of SNPs used in each model is represented. Since, they are five type of receptors and two measurements, ten different GLM models were estimated. Consequently, there was a different number of selected SNPs for each one.

Table 7. Number of significant SNPs (p-value<0.2) in each measure of log(Expression) and Entropy for each response variables (IGH, IGK, IGL, TRA, TRB)

	IGH	IGK	IGL	TRA	TRB
log(Expression)	72,641	72,681	72,815	72,926	72,555
Entropy	72,191	73,436	72,418	72,867	72,715

The number of SNPs with a p-value < 0.2 was quite similar across all variables. Around 73,000 SNPs were associated with Expression and Entropy of IG and TCR and, as observed in the heatmaps (Figure 14 and Figure 15), introducing these SNPs in the methods allowed the majority of them to get high correlations in the testing sets.

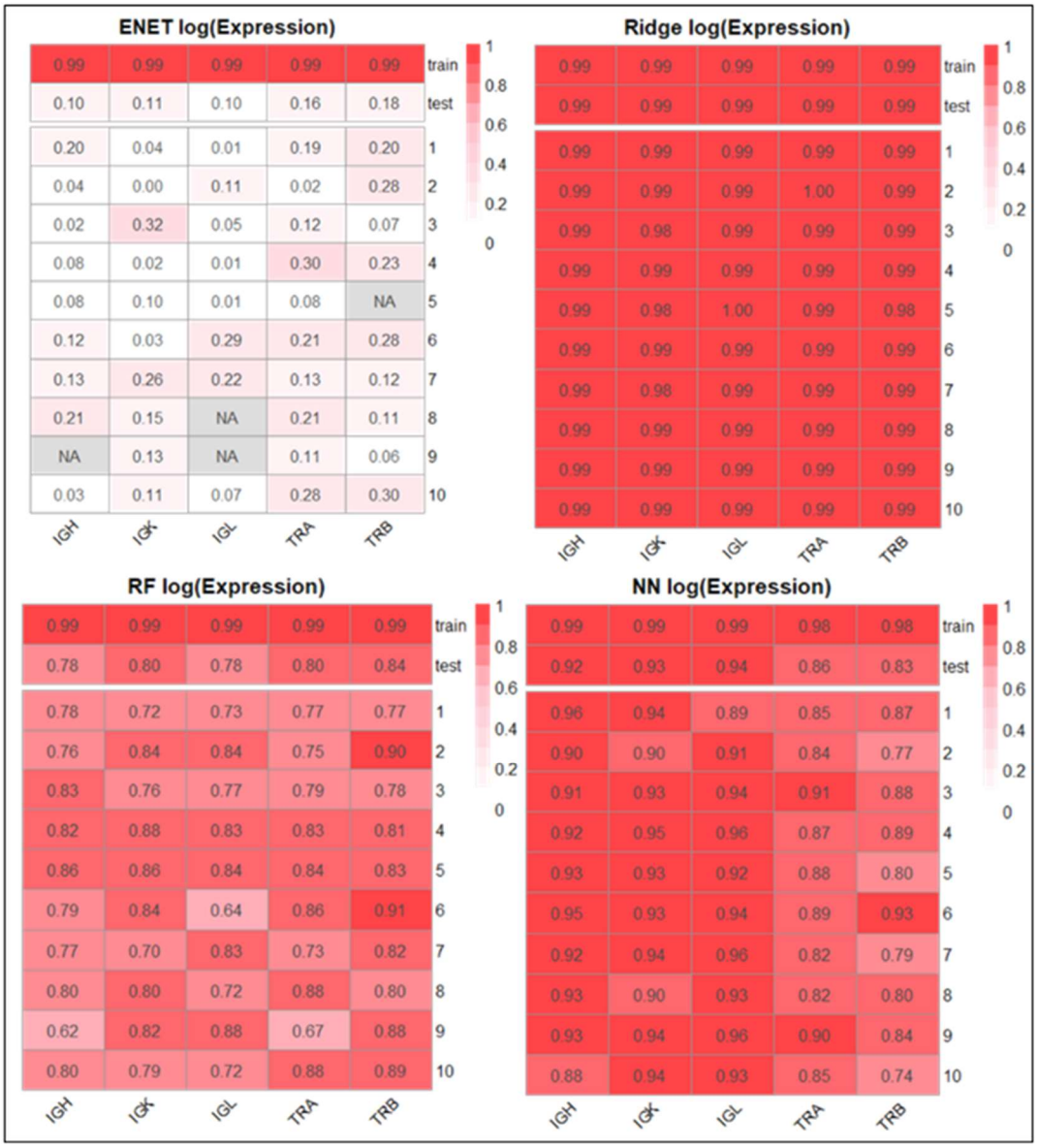

Figure 14. Correlation heatmaps in log(Expression)) in the four methods with GLM filtering in the whole dataset (cut-off = 0.2). Abbreviations: ENET: Elastic Net, RF: Random Forest, NN: Neural Network

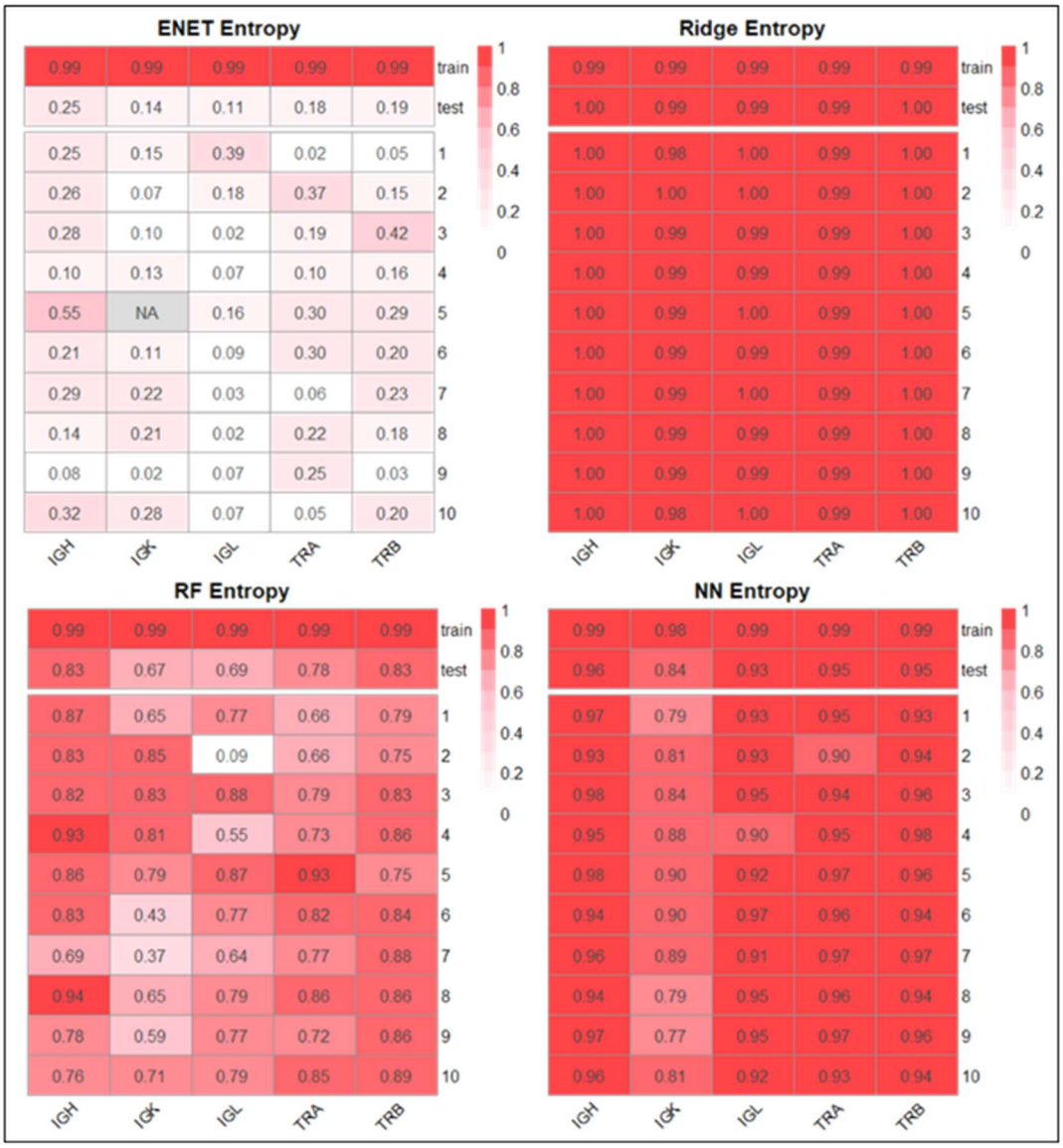

Figure 15. Correlation heatmaps in Entropy in the four methods with GLM filtering in the whole dataset (cut-off = 0.2). Abbreviations: ENET: Elastic Net, RF: Random Forest, NN: Neural Network

In this first scenario, huge differences between ENET correlations in comparison to the other methods were observed. In the two measures (log(Expression) and Entropy) ENET was the method which showed the lowest correlations ($\rho \leq 0.25$) across all the test sets while RiR being also a penalized regression method, presented the highest correlations independently of the sample test used and the variable of interest to predict $(\rho \approx 1)$. NN was positioned in a second place behind RiR regression but showing certain variability within dependent variables, $SD(TRB_{Expression})=0.0591$, $SD(IGK_{Entropy})=0.0505$ (Standard Deviation (SD) tables could be found in Annexes pages 15-18). Related to RF, despite its greatest correlations ($\rho \geq 0.65$), they were lower than the observed in the other two

methods. Moreover, these values varied along the ten testing sets iterations. Proof of this was the prediction calculated for Entropy variables where in IGK and IGL correlation dropped due to the heterogeneity observed among iterations, $SD(IGK_{Entropy})=0.1654$, $SD(IGL_{Entropy})=0.2359$.

Another important point to mention in this scenario is the "NA" observed values in ENET. These missing correlations occurred because the estimation for all the individuals for both, log(Expression) and Entropy was the same. Thus, correlation could not be calculated between observed and predicted values. This problem could be due a lack of convergency in the method.

4.3.2 Scenario 2: GLM p-value < 0.5 in the training dataset

In this second scenario, sample split was done before performing GLM, since the filtering was applied only in the training dataset. Due to this partition, the number of associated SNPs varied across the training samples. Thus, one SNP could be associated with the first training set but removed in the next GLM with the second train set.

Table 8. Median of the number of significant SNPs (p-value<0.5 in the train samples) in each measure of log(Expression) and Entropy for each response variable (IGH, IGK, IGL, TRA, TRB)

	IGH	**IGK**	**IGL**	**TRA**	**TRB**
log(Expression)	182,009	182,016	182,312	182,005	181,742.5
Entropy	181,946.5	183,388	182,412.5	181,958.5	181,604

A higher p-value on the GLM models implied a major number of variables used in the prediction analysis (Table 8). Despite the high dimensionality problem in this project, using a cut-off too restrictive (like in the first scenario) might be an obstacle to our methods: significant SNPs in training sets could be so specific to these samples that they might not be related to testing sets. However, as it will be presented next, results in this scenario were far from be promising:

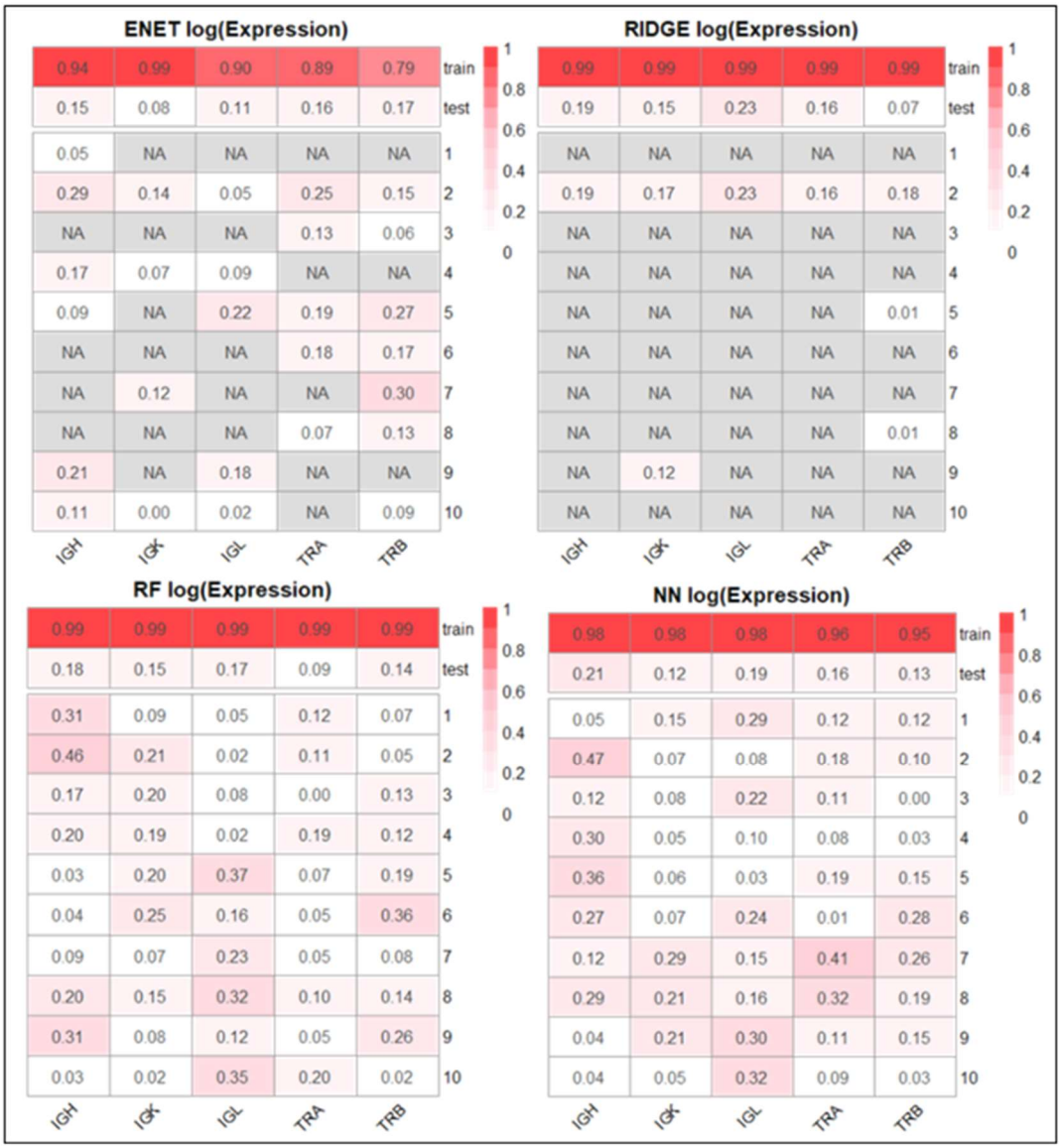

Figure 16. Correlation heatmaps in log(Expression) (A) and Entropy (B) measures in the four methods with GLM filtering in the training dataset (cut-off = 0.5). Abbreviations: ENET: Elastic Net, RF: Random Forest, NN: Neural Network

There was a huge difference from what we have seen in the previous scenario, the correlations (Figure 16, Figure 17) decreased to values minor than 0.3 on testing sets on average for all the methods. Furthermore, in this case, none of the methods showed a good performance. Filtering SNPs with a cut-off equal to 0.5 in the training dataset severely impaired the results. Any of these, methods were incapable of working with such specific training SNPs to predict in the testing.

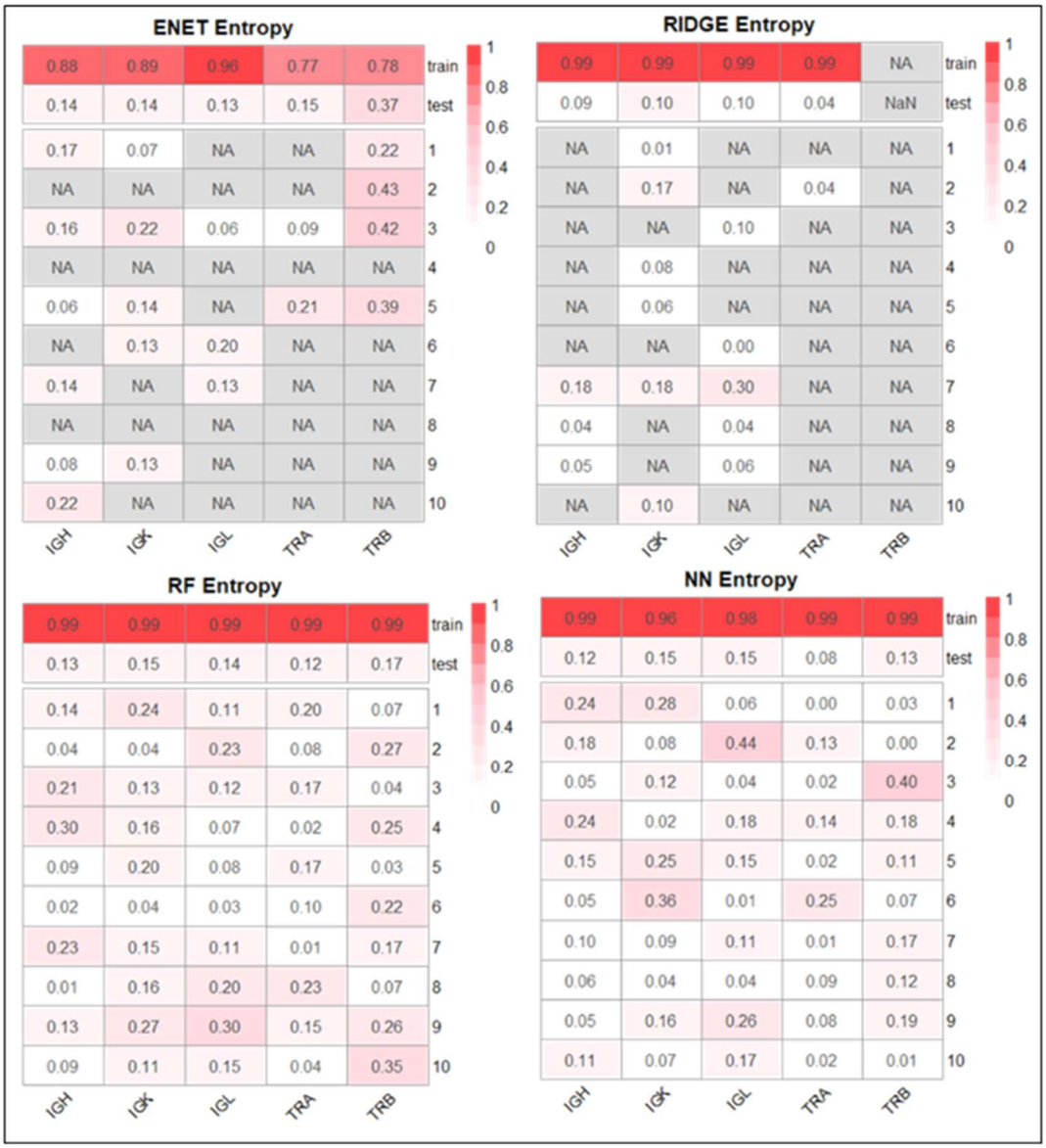

Figure 17. Correlation heatmaps in Entropy in the four methods with GLM filtering in the training dataset (cut-off = 0.5). Abbreviations: ENET: Elastic Net, RF: Random Forest, NN: Neural Network

Related to average correlations, they were calculated avoiding missing values. However, if we focus on TRB Entropy estimated by RiR, there was a convergency problem in the training set which prevented from estimating a correlation in this particular variable.

4.3.3 Scenario 3: GLM p-value < 0.5 in the whole dataset

Based on the results obtained in the two previous scenarios, a final scenario taking into account the high dimensionality and the excess of specific SNPs was considered. Hence, a similar number of SNPs as in the second scenario was chosen to take part in the analysis (Table 9).

Table 9. Number of significant SNPs (p-value<0.5) in each measure of log(Expression) and Entropy for each response variable (IGH, IGK, IGL, TRA, TRB)

	IGH	IGK	IGL	TRA	TRB
Log(Expression)	181,547	181,614	181,998	181,877	181,322
Entropy	181,819	182,701	182,027	181,655	181,431

A cut-off of 0.5 was determined for the last scenario where the GLM filter was applied in the whole dataset. Taking this into account, methods were predicting with a considerable number of SNPs not associated with the dependent variable in the GLM model (around 182,000).

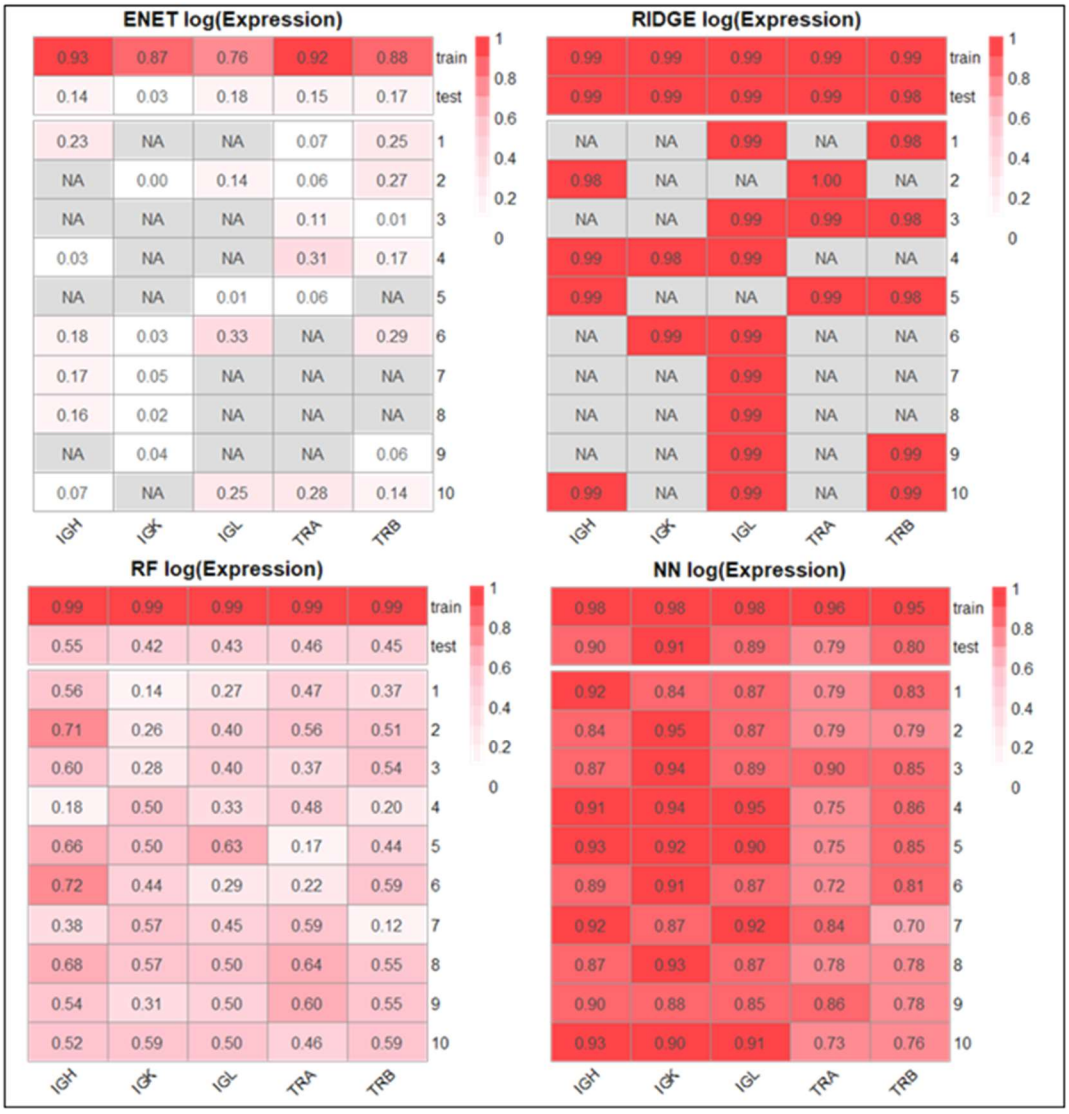

Figure 18. Correlation heatmaps in log(Expression) measures in the four methods with GLM filtering in the whole dataset (cut-off = 0.2). Abbreviations: ENET: Elastic Net, RF: Random Forest, NN: Neural Network

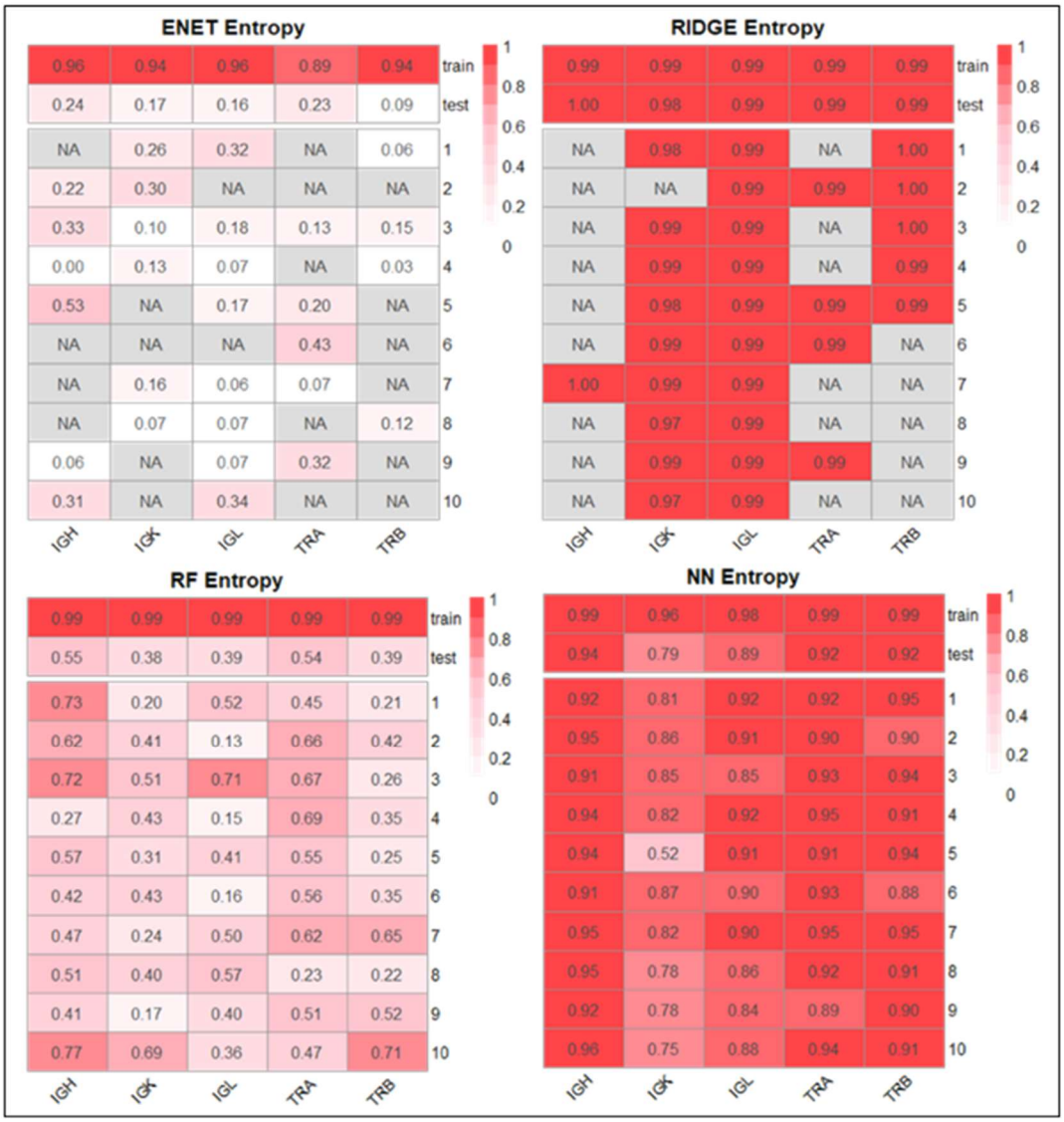

Figure 19. Correlation heatmaps in Entropy in the four methods with GLM filtering in the whole dataset (cut-off = 0.2). Abbreviations: ENET: Elastic Net, RF: Random Forest, NN: Neural Network

Through heatmaps (Figure 18 and Figure 19), it is shown that for both, log(Expression) and Entropy correlations, NN was the method showing the highest correlation values and the less variable across the ten iterations. In comparison with RiR ($\rho(mean) = 0.99$) and despite its correlations were higher than NN on average ($\rho(mean) \geq 0.79$), there was a great number of missing values which meaning that RiR fault to converge. Related to RF results, the range of correlation values was so wide that on average, $SD(RF_{Expression})>0.11$, $SD(RF_{Entropy})>0.13$, predicting testing results were bound to 0.5.

Finally, focusing on NN, predictions in log(Expression) and Entropy were very similar across the response variables: TRA and TRB log(Expression) had lower correlations in comparison with the correlations obtained in Entropy, although all of them were

extraordinary ($\rho > 0.7$). On the contrary, correlations in IGK were higher in log(Expression) $\left(\overline{\rho} = 0.96,\ \text{SD} = 0.0364\right)$ than in Entropy where it was observed more variability than with the other receptors $\left(\overline{\rho} = 0.79,\ \text{SD} = 0.1011\right)$.

5. DISCUSSION

Lately, the use of machine learning techniques is widely used to build predictive models in the context of high-throughput omics data in cancer research[42]. In this study, we applied different machine learning methods to find the best prediction accuracy using genetic data to predict immune infiltration in PC. We found that there was a large number of SNPs associated with Expression and Entropy variables showing a distinct genetic signature associated with the tumoral immune infiltration in PC. In addition, we observed that a different selection of significant SNPs was associated withing each measure suggesting that they measured different characteristics of the immune repertoire and they were genetically different explained. Furthermore, we demonstrated that NN is the technique showing the better prediction accuracy and also the one that is more consistent among the different testing sets. Choosing NN as the optimal prediction method in scenario 3, accurate results were obtained for all the measures showing better results for IG Expression than for TCR, while in Entropy, TCR correlations were higher than IGK and IGL, the latter being the worst predicted outcome.

The immune infiltration associated with germline variants is a growing field and our results were consistently with the ones found by Shahamatdar et al. who identified multiple germline genetic features associated with tumor immune-phenotypes[43]. Therefore, to predict immune feature characteristics such as richness and diversity is possible using SNPs values as we have shown in this master thesis when the appropriate methodology is considered.

Genetic data analyses are characterized by the composition of a huge number of variables in comparison with the small sample size of the participants involved in the studies. Usually, SNPs are the typical variables used to study genetics although they have a specific biological structure which requires a special methodological treatment.

Hence, in this project, several methodological properties were explored before performing the prediction analysis: PS which might be a bias in genetic studies was solved performing a PCA to consider a sample filter in which only Caucasian individuals were chosen. Therefore, the race of the individuals did not influence in rest of the analysis. Additionally, rare genetic variants (MAF < 0.05) and the highly correlated ones (LD > 0.9) were excluded for the analysis to avoid methodological problems in the algorithm convergencies.

Despite all these biological and methodological aspects considered, high dimensionality is still a drawback in this study. Machine learning methods are often chosen as the best option to analyze high-throughput data. However, a feature selection approach is needed to solve the high dimensionality problem and improve the prediction accuracy of the models.

In this regards, several authors[44–47] propose the use of univariate GLM as an acceptable filter to reduce the large number of variables. Nevertheless, the cut-off point and the dataset selected to apply this filter affects considerably to the final prediction performance. A very restrictive cut-off in the whole dataset (as shown in scenario 1) implies overfitting in all methods. This implied that the predictions could not be externally validated since some of the selected SNPs included in the training set may be no relevant in the completely independent dataset.

However, when the GLM was applied to filter SNPs only in the training set (scenario 2), the prediction accuracy decreased in all methods despite the use of a no restrictive cut-off, probably due to the lack of significant SNPs in the testing dataset. Therefore, the dataset used for the prediction model was too specific to the training dataset.

On the other hand, results improved when the cut-off was less restrictive (scenario 3) and it was applied to the whole dataset. In such a case, the methods tend to show high and consistent correlations with no apparent overfitting since not only significant SNPs associated with the whole dataset are introduced, but also a reasonable number of "noise" SNPs. The results found in this scenario are promising. In spite of the considerable number of insignificant SNPs, NN prediction ability has not been affected whereas RF and RiR got worse correlations and consistency respectively, as it was already mentioned. Two reasons may explain this situation. The first one, the internal bootstrap process associated with RF that could include additional not essential SNPs. The second one, the

huge number of SNPs in this scenario, makes difficult the convergence of RiR model. However, feature selection approach remained unsolved and further research on this regard is needed.

The use of GLM as filtering has many limitations, in one hand, the multivariate structure of the data was not considered and on the other hand, the correlation structure among the SNPs, which might be a key point in prediction, was completely destroyed with the univariate model.

Having all this considered, feature selection is a key point of study in predictive analysis and more techniques must be further investigated to solve GLM limitations regarding genetic data. For example, graph domination[48] is presented as a reliable variable selection method that considers correlation patterns among SNPs through a graph-theory approach.

But these limitations are not observed equally among the three methods tested. Indeed, there has been observed some differences among prediction methods, specially between ENET and RiR. Both penalized regression methods presented a really distinct correlation accuracy due to their mathematical foundation. RiR (like RF and NN) uses all variables in the analysis while ENET performs a previous variable selection that, as it has shown in the results, severely impairs the correlations. Regardless the different scenarios, ENET always presented a huge variability among the testing sets and, in both methods, consistency problems were observed (missing correlations arise when the same predicted value is assigned to all individuals in the sample).

In RF and NN, although missing values do not appear, there were more heterogeneous correlations in RF than in NN (for all measures and variables). If we focus on the third scenario, the mean value of the correlations from the testing dataset were representative of the ones estimated along the 10 iterations and, taking into account the large number of variables introduced in this scenario, NN worked exceptionally well with correlations over 80% in both Expression and Entropy measures. This could be due to the well-known lack of predictive accuracy of the RF and the regression models.

Despite the limitations mentioned above (overfitting, sparsity and multicollinearity) studied also in other surveys[49], machine learning algorithms are capable of mitigating these problems since they are used without any assumptions (such as normality) and, most

importantly, they are not assessed using a p-value which might be a problem in these large studies[50].

In summary, SNPs, which we have shown to be associated with the tumoral immune infiltration in PC, are complex variables that should be treated according with its biological structure and solving the methodological problems discussed here. In this project, we have proposed an accurate predictive model to integrate genetic data with immune infiltration using different filtering functions that may deal with the main drawbacks. Finally, we have shown that the use of NN in a specific scenario overcome the overfitting and over-specificity problems. Being able to predict the immune infiltration with genetic variants will allow us to decipher new biological insights extremely necessary in PC research.

6. CONCLUSIONS

In view of the results obtained with the different methods across the three scenarios, the following conclusions can be drawn:

- There was a large number of significant SNPs associated with Expression and Entropy suggesting that the tumoral immune infiltration may be modulated by genetic susceptibility in PC. Due to the short number of SNPs overlapping between both measures, they should be considered separately.
- NN presented better prediction results based on Pearson correlation than ENET and RF and a better consistency among the testing sets than RiR although its performance depends on the feature selection criteria.
- Scenario 3: GLM p-value < 0.5 filter showed to be the most properly scenario according to the overfitting and feature selection problem.
- Secondary scenarios showed that further research is needed to explore other feature selection approaches and a completely independent testing dataset.

7. NEXT STEPS

Results obtained in this project have demonstrated the different accuracy predictions obtained in each scenario. Because of that, one of the main issues to study in a future plan will be to explore new variable selection approaches. Currently, there are some lines of investigation[51,52] focus on feature selection methodologies which will be of great interest to explore in the context of this project. Furthermore, given the special nature of genetic data, it would be convenient to consider some own properties of these variables, such as their genetic architecture.

Correlation structure is a key point to take into account when SNPs are investigated. Despite, LD is a well-known genetic filter, a further exploration is needed to include this characteristic in the previous selection filter. Then, having these pre-processing steps done, the predictive methods will be applied and results validated in a completely independent testing set.

One next step of this project is to predict the immune infiltration on a completely independent dataset in order to find new exposure factors affecting the development of PC. To this end, PanGenEU[27] will be used. This database is a case-control study with a large European population (2,500 PC cases and 1,600 controls) that collected epidemiological and clinical information. The study also obtained biosamples (blood, saliva, urine, toenails and tissue), which have already been used to profile genomics, epigenomics, metabolomics, and microbiome. To evaluate the prediction accuracy, the genetic variants measured in the PanGenEU will be used. The model extracted from the scenario 3 using NN will be applied since it has been the method which has better fit the data and whose predictions have outperformed the other. With these predictions, the association between the main risk factors (tobacco, diet, allergies, among others) and PC immunological infiltration will be assessed.

Another pending objective to explore is to combine all Expression and Entropy variables either in two unique measures to determine richness and diversity levels or considering IG and TCR measures as a global measure. Methods that allow us to work with multiple dependent variables will be assessed in combination with the methods employed in this work.

Finally, simulated study will be carried out to assess the performance of the four methods considered in this master thesis under different prefixed conditions.

8. BIBLIOGRAPHY

1. Lepage C, Capocaccia R, Hackl M, Lemmens V, Molina E, Pierannunzio D, et al. Survival in patients with primary liver cancer, gallbladder and extrahepatic biliary tract cancer and pancreatic cancer in Europe 1999-2007: Results of EUROCARE-5. Eur J Cancer. 2015;51(15):2169–78.

2. Aune D, Greenwood DC, Chan DSM, Vieira R, Vieira AR, Navarro Rosenblatt DA, et al. Body mass index, abdominal fatness and pancreatic cancer risk: A systematic review and non-linear dose-response meta-analysis of prospective studies. Ann Oncol. 2012;23(4):843–52.

3. Molina-Montes E, van Hoogstraten L, Gomez-Rubio P, Löhr M, Sharp L, Molero X, et al. Pancreatic cancer risk in relation to lifetime smoking patterns, tobacco type, and dose-response relationships. Cancer Epidemiol Biomarkers Prev. 2020;29(5):1009–18.

4. Bosetti C, Rosato V, Li D, Silverman D, Petersen GM, Bracci PM, et al. Diabetes, antidiabetic medications, and pancreatic cancer risk: an analysis from the International Pancreatic Cancer Case-Control Consortium. Ann Oncol. 2014;25(10):2065–72.

5. Lucenteforte E, La Vecchia C, Silverman D, Petersen GM, Bracci PM, Ji BT, et al. Alcohol consumption and pancreatic cancer: A pooled analysis in the International Pancreatic Cancer Case-Control Consortium (PanC4). Ann Oncol. 2012;23(2):374–82.

6. Kirkegård J, Mortensen FV, Cronin-Fenton D. Chronic Pancreatitis and Pancreatic Cancer Risk: A Systematic Review and Meta-analysis. Off J Am Coll Gastroenterol | ACG. 2017;112(9).

7. Wolpin BM, Chan AT, Hartge P, Chanock SJ, Kraft P, Hunter DJ, et al. ABO blood group and the risk of pancreatic cancer. J Natl Cancer Inst. 2009;101(6):424–

31.

8. Klein AP. Genetic susceptibility to pancreatic cancer. Mol Carcinog. 2012;51(1):14–24.

9. Gomez-Rubio P, Zock J-P, Rava M, Marquez M, Sharp L, Hidalgo M, et al. Reduced risk of pancreatic cancer associated with asthma and nasal allergies. Gut. 2017 Feb 1;66(2):314 LP – 322.

10. López de Maturana E, Rodríguez JA, Alonso L, Lao O, Molina-Montes E, Martín-Antoniano IA, et al. A multilayered post-GWAS assessment on genetic susceptibility to pancreatic cancer. Genome Med. 2021;13(1):1–18.

11. Sayaman RW, Saad M, Thorsson V, Hu D, Hendrickx W, Roelands J, et al. Germline genetic contribution to the immune landscape of cancer. Immunity. 2021;54(2):367-386.e8.

12. Barnes TA, Amir E. HYPE or HOPE: The prognostic value of infiltrating immune cells in cancer. Br J Cancer. 2017;117(4):451–60.

13. Foucher ED, Ghigo C, Chouaib S, Galon J, Iovanna J, Olive D. Pancreatic ductal adenocarcinoma: A strong imbalance of good and bad immunological cops in the tumor microenvironment. Front Immunol. 2018;9(MAY):1–8.

14. Ino Y, Yamazaki-Itoh R, Shimada K, Iwasaki M, Kosuge T, Kanai Y, et al. Immune cell infiltration as an indicator of the immune microenvironment of pancreatic cancer. Br J Cancer. 2013;108(4):914–23.

15. Martinez-Bosch N, Vinaixa J, Navarro P. Immune evasion in pancreatic cancer: From mechanisms to therapy. Cancers (Basel). 2018;10(1):1–16.

16. Janeway CA Jr, Travers P, Walport M et al. Immunobiology: The Immune System in Health and Disease. 2001

17. Wouters MCA, Nelson BH. Prognostic significance of tumor-infiltrating B cells and plasma cells in human cancer. Clin Cancer Res. 2018;24(24):6125–35.

18. Pineda S, Lopez de Maturana E, Yu K, Ravoor A, Wood I, Malats N, et al. Landscape of Tumor-Infiltrating B and T cell Repertoire in Pancreatic Cancer. Front Immunol. 2021;

19. Liston A, Goris A. The origins of diversity in human immunity. Nat Immunol. 2018 Mar;19(3):209–10.

20. Definition of GWAS - NCI Dictionary of Genetics Terms - National Cancer Institute. Available from: https://www.cancer.gov/publications/dictionaries/genetics-dictionary/def/gwas

21. Jehan T, Lakhanpaul S. Single nucleotide polymorphism (SNP) - methods and applications in plant genetics: A review. Indian J Biotechnol. 2006;5(4):435–59.

22. Mao X, Young BD, Lu Y-J. The application of single nucleotide polymorphism microarrays in cancer research. Curr

23. Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D. Benefits and limitations of genome-wide association studies. Nat Rev Genet

24. Hellwege JN, Keaton JM, Giri A, Gao X, Velez Edwards DR, Edwards TL. Population Stratification in Genetic Association Studies. Curr Protoc Hum Genet. 2017;95(1):1.22.1-1.22.23.

25. Slatkin M. Linkage disequilibrium - Understanding the evolutionary past and mapping the medical future. Nat Rev Genet. 2008;9(6):477–85.

26. The Cancer Genome Atlas Program - National Cancer Institute Available from: https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga

27. Gomez-Rubio P, Piñero J, Molina-Montes E, Gutiérrez-Sacristán A, Marquez M, Rava M, et al. Pancreatic cancer and autoimmune diseases: An association sustained by computational and epidemiological case–control approaches. Int J Cancer. 2019;144(7):1540–9.

28. Bolotin DA, Poslavsky S, Mitrophanov I, Shugay M, Mamedov IZ, Putintseva E V., et al. MiXCR: Software for comprehensive adaptive immunity profiling. Nat Methods. 2015;12(5):380–1.

29. Gauch H, Qian S, Piepho H-P, Zhou L, Chen R. Effective principal components analysis of SNP data. bioRxiv. 2018;393611.

30. Hoerl AE, Kennard RW. Ridge Regression: Applications to Nonorthogonal

Problems. Technometrics. 1970;12(1):69–82.

31. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B Stat Methodol. 2005;67(2):301–20.

32. Breiman L. Random forests. Machine Learning. 2001;45:5–32.

33. Breiman L. Bagging predictors. Machine Learning. 1996;24:123–40.

34. Dietterich TG. An experimetnal comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. Machine Learning. 2000;40:139–57.

35. Ho TK. The random subspace method for constructing decision forests. IEEE Trans Pattern Anal Mach Intell. 1998;20(8):832–44.

36. Hastie T, Tibshirani R, Friedman J. The elements of Statistical Learning. Stanford, Calfiornia: Springer; 2008.

37. Bell J. Artificial Neural Networks. In: Machine Learning: Hands-On for Developers and Technical Professionals. 2014. p. 91–116.

38. Martín Martín Q. Investigación Operativa. Prentice Hall; 2005.

39. Sabroso Lasa S. Uso de Técnicas Semánticas y Aprendizaje para el Desarrollo de Sistemas de Recomendación basados en Contenido. Universidad de Zaragoza; 2020.

40. Tibshirani R. Regression Shrinkage and Selection Via the Lasso. J R Stat Soc Ser B

41. R: The R Project for Statistical Computing Available from: https://www.r-project.org/

42. Elkhader J, Elemento O. Artificial intelligence in oncology: From bench to clinic. Semin Cancer Biol

43. Shahamatdar S, He MX, Reyna MA, Gusev A, AlDubayan SH, Van Allen EM, et al. Germline Features Associated with Immune Infiltration in Solid Tumors. Cell Rep

44. Le Floch É, Guillemot V, Frouin V, Pinel P, Lalanne C, Trinchera L, et al.

Significant correlation between a set of genetic polymorphisms and a functional brain network revealed by feature selection and sparse Partial Least Squares. Neuroimage. 2012;63(1):11–24.

45. Kooperberg C, LeBlanc M, Obenchain V. Risk prediction using genome-wide association studies. Genet Epidemiol. 2010;34(7):643–52.

46. Arabnejad M, Montgomery CG, Gaffney PM, McKinney BA. Nearest-Neighbor Projected Distance Regression for Epistasis Detection in GWAS With Population Structure Correction. Front Genet. 2020;11(July):1–8.

47. Seral-Cortes M, Sabroso-Lasa S, De Miguel-Etayo P, Gonzalez-Gross M, Gesteiro E, Molina-Hidalgo C, et al. Development of a Genetic Risk Score to predict the risk of overweight and obesity in European adolescents from the HELENA study. Sci Rep. 2021;11(1):1–11.

48. Sun S, Miao Z, Ratcliffe B, Campbell P, Pasch B, El-Kassaby YA, et al. SNP variable selection by generalized graph domination. PLoS One. 2019;14(1):1–18.

49. Xu C, Jackson SA. Machine learning and complex biological data The revolution of biological techniques and demands for new data mining methods. 2019;1–4.

50. Pineda S, Sirota M. Determining Significance in the New Era for P Values. J Pediatr Gastroenterol Nutr. 2018;67(5):547–8.

51. Cai J, Luo J, Wang S, Yang S. Feature selection in machine learning: A new perspective. Neurocomputing. 2018;300:70–9.

52. 20 Recursive Feature Elimination | The caret Package Available from: https://topepo.github.io/caret/recursive-feature-elimination.html

PREDICTIVE ANALYSIS TO FIND GERMLINE GENETIC SUSCEPTIBILITY ASSOCIATED WITH THE TUMORAL IMMUNE INFILTRATION IN PANCREATIC CANCER

Supplementary Material

INDEX OF ANNEXES

1. PEARSON RESIDUALS

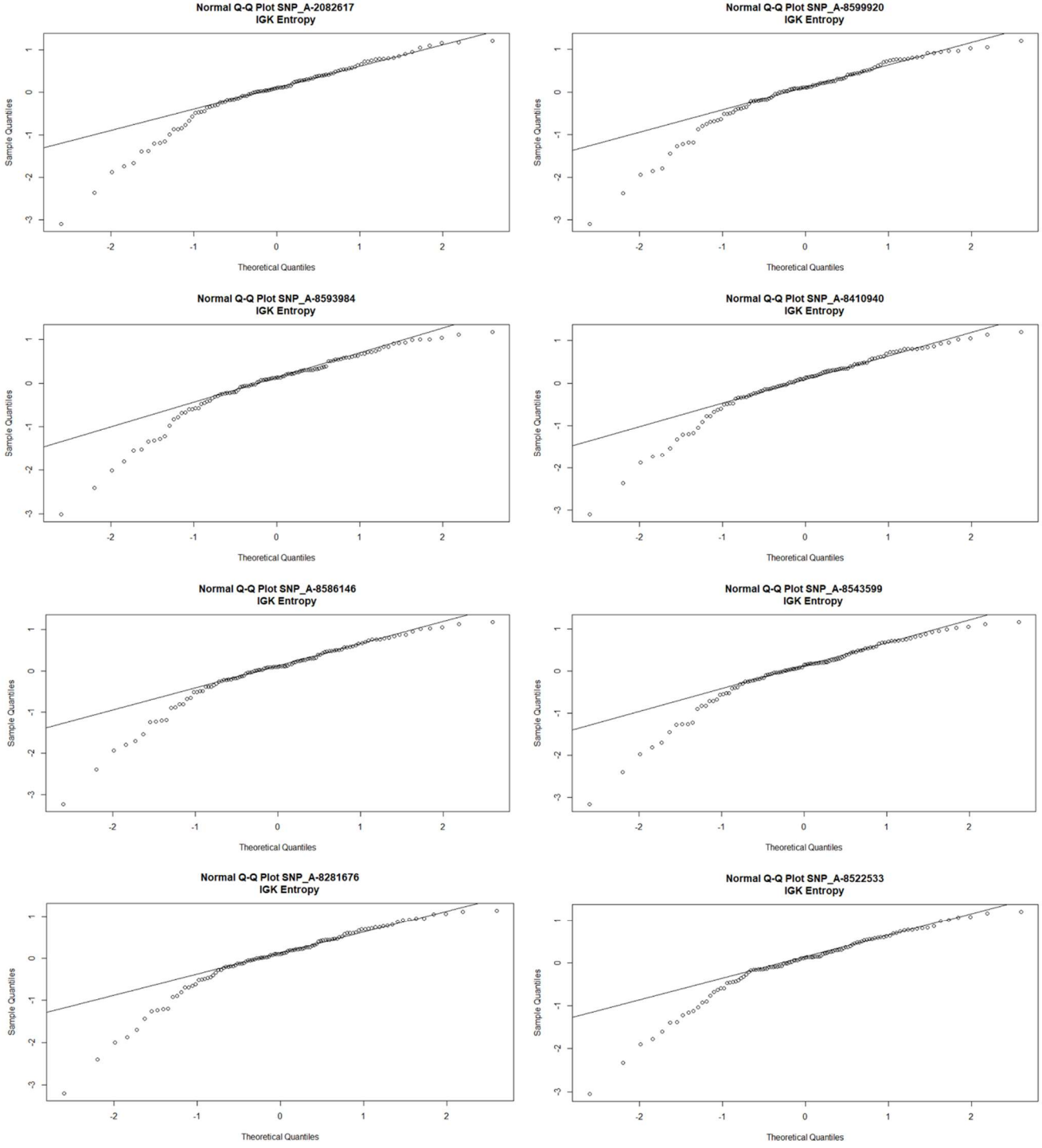

Figure 1. *Pearson residuals QQ-plots in IGK (8 random SNPs)*

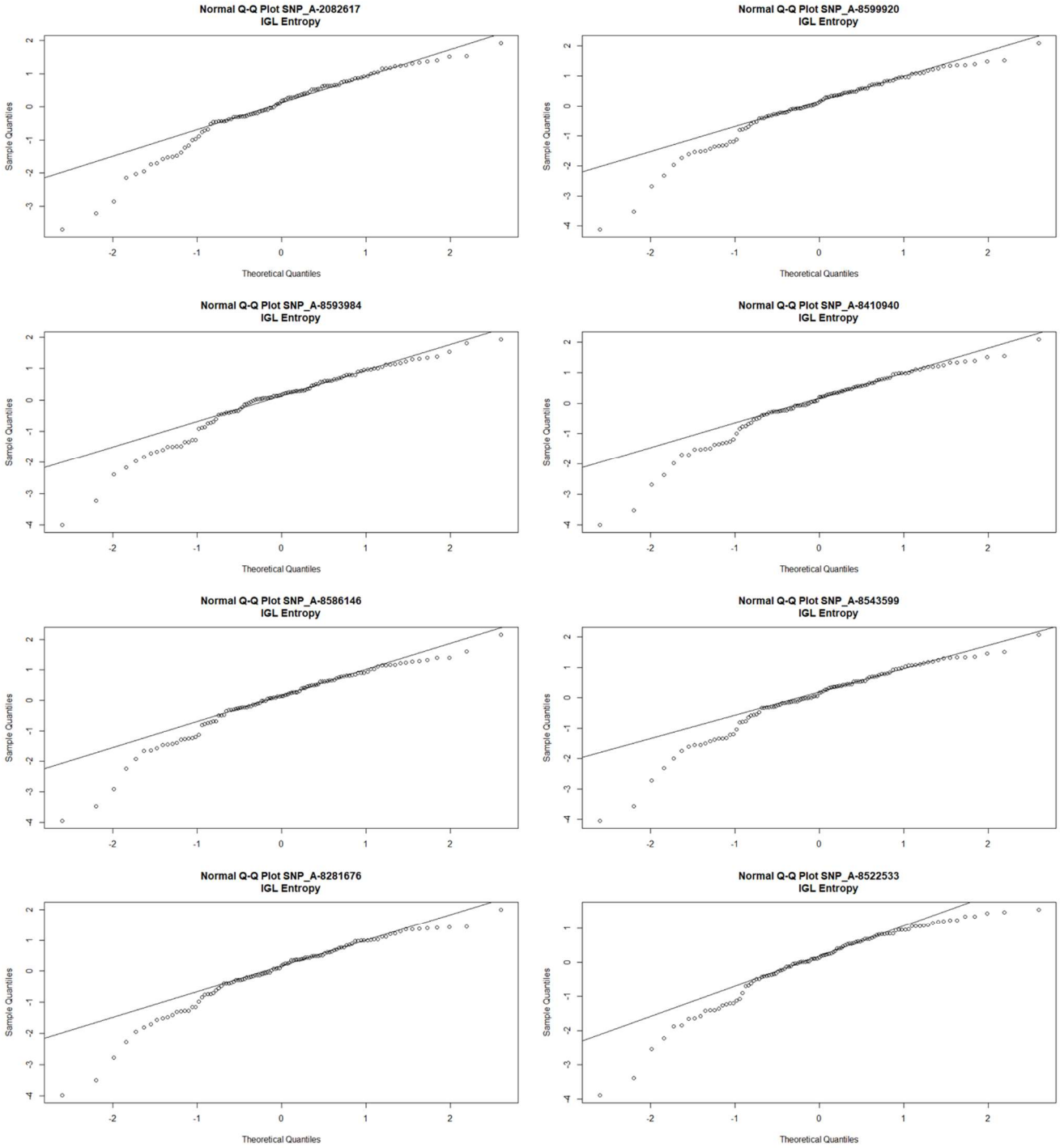

Figure 2. *Pearson residuals QQ-plots in IGL (8 random SNPs)*

2. PEARSON CORRELATION TABLES

2.1 SCENARIO 1

Table 1. *Expression Train Pearson Correlations with ENET in Scenario 1*

Expression Train	1	2	3	4	5	6	7	8	9	10
IGH	1	1	1	1	1	1	1	1	NA	1
IGK	1	1	1	1	0.9325	1	1	1	1	1
IGL	1	1	1	1	1	1	1	NA	NA	1
TRA	1	1	1	1	0.9203	1	1	1	1	1
TRB	1	1	1	1	NA	1	1	1	0.5849	1

Table 2. *Expression Test Pearson Correlations with ENET in Scenario 1*

Expression Test	1	2	3	4	5	6	7	8	9	10
IGH	0.1980	0.0378	0.0228	0.0848	0.0801	0.1199	0.1295	0.2072	NA	0.0319
IGK	0.0362	0.0024	0.3207	0.0201	0.0985	0.0263	0.2582	0.1539	0.1277	0.1059
IGL	0.0076	0.1106	0.0540	0.0086	0.0080	0.2916	0.2226	NA	NA	0.0663
TRA	0.1947	0.0204	0.1152	0.3015	0.0789	0.2059	0.1260	0.2076	0.1126	0.2794
TRB	0.2037	0.2761	0.0662	0.2333	NA	0.2783	0.1193	0.1114	0.0565	0.2952

Table 3. *Entropy Train Pearson Correlations with ENET in Scenario 1*

Entropy Train	1	2	3	4	5	6	7	8	9	10
IGH	1	1	1	1	1	1	1	0.9995	1	1
IGK	1	1	1	1	NA	1	1	1	1	1
IGL	1	1	1	1	1	1	1	1	1	1
TRA	1	0.8971	1	1	0.9203	1	1	1	1	0.9999
TRB	0.9225	1	1	0.9683	1	1	0.9905	1	0.9697	1

Table 4. *Expression Test Pearson Correlations with ENET in Scenario 1*

Entropy Test	1	2	3	4	5	6	7	8	9	10
IGH	0.2500	0.2632	0.2788	0.1017	0.5486	0.2055	0.2868	0.1414	0.0774	0.3211
IGK	0.1545	0.0682	0.0967	0.1334	NA	0.1121	0.2164	0.2102	0.0235	0.2751
IGL	0.3925	0.1827	0.0181	0.0684	0.1622	0.0881	0.0286	0.0230	0.0699	0.0705
TRA	0.0193	0.3662	0.1929	0.1009	0.2974	0.2993	0.0574	0.2154	0.2544	0.0464
TRB	0.0523	0.1512	0.4235	0.1553	0.2875	0.2005	0.2308	0.1844	0.0286	0.2031

Table 5. *Expression Train Pearson Correlations with RiR in Scenario 1*

Expression Train	1	2	3	4	5	6	7	8	9	10
IGH	0.9997	0.9997	1	1	0.9996	1	0.9997	1	1	1
IGK	0.9996	1	1	0.9996	1	1	0.9996	1	1	1
IGL	0.9997	0.9997	1	1	0.9996	1	1	0.9997	0.9997	0.9996
TRA	0.9997	1	0.9997	0.9998	0.9997	0.9997	0.9998	0.9997	0.9997	0.9997
TRB	0.9997	0.9996	1	0.9997	0.9998	0.9996	0.9998	1	0.9996	1

Table 6. *Expression Test Pearson Correlations with RiR in Scenario 1*

Expression Test	1	2	3	4	5	6	7	8	9	10
IGH	0.9886	0.9873	0.9918	0.9910	0.9893	0.9910	0.9897	0.9865	0.9941	0.9896
IGK	0.9862	0.9884	0.9841	0.9874	0.9833	0.9878	0.9808	0.9864	0.9901	0.9864
IGL	0.9910	0.9884	0.9930	0.9912	0.9951	0.9936	0.9909	0.9910	0.9929	0.9945
TRA	0.9918	0.9961	0.9891	0.9888	0.9947	0.9920	0.9860	0.9917	0.9880	0.9915
TRB	0.9877	0.9929	0.9871	0.9863	0.9835	0.9905	0.9899	0.9896	0.9900	0.9911

Table 7. *Entropy Train Pearson Correlations with RiR in Scenario 1*

Entropy Train	1	2	3	4	5	6	7	8	9	10
IGH	1	0.9999	1	0.9998	0.9998	0.9999	0.9999	0.9998	0.9998	0.9999
IGK	1	1	0.9995	1	1	1	1	1	1	0.9995
IGL	0.9998	0.9998	1	0.9998	1	0.9998	0.9998	1	0.9999	0.9998
TRA	0.9997	1	1	1	0.9997	0.9996	0.9997	1	0.9996	0.9997
TRB	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	1	0.9999	0.9999

Table 8. *Entropy Test Pearson Correlations with RiR in Scenario 1*

Entropy Test	1	2	3	4	5	6	7	8	9	10
IGH	0.9970	0.9955	0.9971	0.9981	0.9979	0.9969	0.9980	0.9978	0.9973	0.9971
IGK	0.9846	0.9955	0.9887	0.9889	0.9874	0.9949	0.9934	0.9859	0.9917	0.9825
IGL	0.9959	0.9951	0.9947	0.9944	0.9972	0.9934	0.9962	0.9942	0.9921	0.9954
TRA	0.9913	0.9900	0.9899	0.9919	0.9912	0.9917	0.9895	0.9888	0.9896	0.9900
TRB	0.9979	0.9979	0.9978	0.9968	0.9969	0.9965	0.9974	0.9975	0.9975	0.9961

Table 9. *Expression Train Pearson Correlations with RF in Scenario 1*

Expression Train	1	2	3	4	5	6	7	8	9	10
IGH	0.9980	0.9981	0.9983	0.9983	0.9981	0.9983	0.9982	0.9981	0.9977	0.9982
IGK	0.9971	0.9982	0.9982	0.9978	0.9981	0.9981	0.9986	0.9987	0.9982	0.9981
IGL	0.9976	0.9970	0.9984	0.9981	0.9984	0.9972	0.9985	0.9977	0.9982	0.9968
TRA	0.9977	0.9967	0.9975	0.9968	0.9975	0.9972	0.9972	0.9971	0.9979	0.9978
TRB	0.9970	0.9970	0.9969	0.9970	0.9974	0.9975	0.9964	0.9970	0.9982	0.9982

Table 10. *Expression Test Pearson Correlations with RF in Scenario 1*

Expression Test	1	2	3	4	5	6	7	8	9	10
IGH	0.7795	0.7566	0.8315	0.8153	0.8594	0.7944	0.7656	0.7989	0.6240	0.7979
IGK	0.7226	0.8374	0.7585	0.8779	0.8558	0.8387	0.7012	0.8040	0.8222	0.7943
IGL	0.7322	0.8365	0.7659	0.8321	0.8438	0.6438	0.8258	0.7224	0.8756	0.7231
TRA	0.7737	0.7494	0.7917	0.8293	0.8398	0.8608	0.7280	0.8819	0.6663	0.8808
TRB	0.7660	0.9046	0.7820	0.8149	0.8287	0.9070	0.8187	0.7976	0.8847	0.8881

Table 11. *Entropy Train Pearson Correlations with RF in Scenario 1*

Entropy Train	1	2	3	4	5	6	7	8	9	10
IGH	0.9971	0.9972	0.9967	0.9978	0.9974	0.9987	0.9984	0.9983	0.9984	0.9978
IGK	0.9958	0.9965	0.9977	0.9972	0.9978	0.9971	0.9966	0.9975	0.9973	0.9981
IGL	0.9964	0.9968	0.9964	0.9965	0.9970	0.9984	0.9956	0.9985	0.9963	0.9957
TRA	0.9976	0.9973	0.9983	0.9984	0.9986	0.9974	0.9978	0.9980	0.9983	0.9970
TRB	0.9975	0.9977	0.9978	0.9974	0.9978	0.9977	0.9979	0.9973	0.9981	0.9971

Table 12. *Entropy Test Pearson Correlations with RF in Scenario 1*

Entropy Test	1	2	3	4	5	6	7	8	9	10
IGH	0.8747	0.8313	0.8171	0.9306	0.8590	0.8313	0.6881	0.9410	0.7770	0.7647
IGK	0.6533	0.8453	0.8266	0.8113	0.7870	0.4334	0.3663	0.6494	0.5885	0.7125
IGL	0.7688	0.0857	0.8846	0.5460	0.8748	0.7698	0.6365	0.7937	0.7693	0.7924
TRA	0.6580	0.6595	0.7904	0.7255	0.9261	0.8243	0.7664	0.8602	0.7157	0.8491
TRB	0.7863	0.7497	0.8311	0.8550	0.7487	0.8392	0.8812	0.8590	0.8611	0.8924

Table 13. *Expression Train Pearson Correlations with NN in Scenario 1*

Expression Train	1	2	3	4	5	6	7	8	9	10
IGH	0.9940	0.9958	0.9906	0.9837	0.9826	0.9958	0.9870	0.9888	0.9826	0.9918
IGK	0.9948	0.9900	0.9931	0.9903	0.9880	0.9863	0.9881	0.9909	0.9914	0.9865
IGL	0.9933	0.9746	0.9896	0.9924	0.9938	0.9895	0.9910	0.9949	0.9950	0.9922
TRA	0.9917	0.9816	0.9814	0.9750	0.9832	0.9811	0.9823	0.9889	0.9903	0.9839
TRB	0.9853	0.9874	0.9783	0.9764	0.9656	0.9798	0.9620	0.9874	0.9849	0.9767

Table 14. *Expression Test Pearson Correlations with NN in Scenario 1*

Expression Test	1	2	3	4	5	6	7	8	9	10
IGH	0.9589	0.9021	0.9066	0.9239	0.9313	0.9478	0.9208	0.9304	0.9271	0.8777
IGK	0.9390	0.8988	0.9322	0.9469	0.9255	0.9282	0.9374	0.8966	0.9419	0.9421
IGL	0.8941	0.9068	0.9369	0.9618	0.9218	0.9425	0.9555	0.9346	0.9637	0.9324
TRA	0.8544	0.8359	0.9122	0.8668	0.8833	0.8898	0.8168	0.8237	0.9014	0.8510
TRB	0.8719	0.7699	0.8761	0.8897	0.7986	0.9264	0.7914	0.8047	0.8371	0.7408

Table 15. *Entropy Train Pearson Correlations with NN in Scenario 1*

Entropy Train	1	2	3	4	5	6	7	8	9	10
IGH	0.9890	0.9901	0.9959	0.9890	0.9938	0.9928	0.9890	0.9945	0.9928	0.9963
IGK	0.9719	0.9761	0.9880	0.9770	0.9728	0.9869	0.9875	0.9856	0.9857	0.9849
IGL	0.9909	0.9803	0.9861	0.9937	0.9848	0.9897	0.9790	0.9884	0.9909	0.9800
TRA	0.9905	0.9959	0.9935	0.9944	0.9889	0.9964	0.9940	0.9953	0.9937	0.9934
TRB	0.9956	0.9942	0.9920	0.9918	0.9939	0.9909	0.9885	0.9920	0.9897	0.9920

Table 16. *Entropy Test Pearson Correlations with NN in Scenario 1*

Entropy Test	1	2	3	4	5	6	7	8	9	10
IGH	0.9726	0.9316	0.9756	0.9517	0.9776	0.9443	0.9611	0.9388	0.9677	0.9609
IGK	0.7877	0.8078	0.8387	0.8801	0.8994	0.8951	0.8934	0.7939	0.7681	0.8098
IGL	0.9324	0.9348	0.9481	0.8971	0.9242	0.9687	0.9150	0.9536	0.9461	0.9230
TRA	0.9476	0.8990	0.9387	0.9493	0.9663	0.9612	0.9694	0.9566	0.9694	0.9283
TRB	0.9273	0.9398	0.9561	0.9756	0.9569	0.9409	0.9740	0.9420	0.9571	0.9427

2.2 SCENARIO 2

Table 17. *Expression Train Pearson Correlations with ENET in Scenario 2*

Expression Train	1	2	3	4	5	6	7	8	9	10
IGH	0.7256	0.9323	NA	0.9987	0.9996	NA	NA	NA	0.9709	0.9961
IGK	NA	1	NA	0.9990	NA	NA	0.9999	NA	NA	0.9921
IGL	NA	1	NA	0.9890	0.9878	NA	NA	NA	0.9680	0.5382
TRA	NA	0.9726	0.9775	NA	0.8433	0.9684	NA	0.6799	NA	NA
TRB	NA	0.8954	0.9843	NA	0.9874	0.6371	0.9268	0.5449	NA	0.5369

Table 18. *Expression Test Pearson Correlations with ENET in Scenario 2*

Expression Test	1	2	3	4	5	6	7	8	9	10
IGH	0.0451	0.2873	NA	0.1668	0.0946	NA	NA	NA	0.2109	0.1062
IGK	NA	0.1425	NA	0.0732	NA	NA	0.1176	NA	NA	0.0006
IGL	NA	0.0495	NA	0.0907	0.2193	NA	NA	NA	0.1770	0.0182
TRA	NA	0.2539	0.1309	NA	0.1893	0.1758	NA	0.0741	NA	NA
TRB	NA	0.1507	0.0552	NA	0.2663	0.1657	0.3004	0.1262	NA	0.0919

Table 19. *Entropy Train Pearson Correlations with ENET in Scenario 2*

Entropy Train	1	2	3	4	5	6	7	8	9	10
IGH	0.9867	NA	0.7795	NA	0.5981	NA	1	NA	0.9442	0.9997
IGK	0.9682	NA	0.9835	NA	0.6272	0.8530	NA	NA	1	NA
IGL	NA	NA	0.9907	NA	NA	0.8973	1	NA	NA	NA
TRA	NA	NA	0.5422	NA	0.9923	NA	NA	NA	NA	NA
TRB	0.5462	0.6778	10.000	NA	0.8919	NA	NA	NA	NA	NA

Table 20. *Entropy Test Pearson Correlations with ENET in Scenario 2*

Entropy Test	1	2	3	4	5	6	7	8	9	10
IGH	0.1680	NA	0.1642	NA	0.0563	NA	0.1369	NA	0.0849	0.2175
IGK	0.0699	NA	0.2163	NA	0.1368	0.1271	NA	NA	0.1250	NA
IGL	NA	NA	0.0578	NA	NA	0.1991	0.1267	NA	NA	NA
TRA	NA	NA	0.0936	NA	0.2052	NA	NA	NA	NA	NA
TRB	0.2216	0.4347	0.4231	NA	0.3903	NA	NA	NA	NA	NA

Table 21. *Expression Train Pearson Correlations with RiR in Scenario 2*

Expression Train	1	2	3	4	5	6	7	8	9	10
IGH	NA	0.9997	NA	NA	NA	NA	NA	NA	NA	NA
IGK	NA	1	NA	NA	NA	NA	NA	NA	0.9998	NA
IGL	NA	0.9997	NA	NA	NA	NA	NA	NA	NA	NA
TRA	NA	1	NA	NA	NA	NA	NA	NA	NA	NA
TRB	NA	1	NA	NA	0.9998	NA	NA	0.9997	NA	NA

Table 22. *Expression Test Pearson Correlations with RiR in Scenario 2*

Expression Test	1	2	3	4	5	6	7	8	9	10
IGH	NA	0.1889	NA	NA	NA	NA	NA	NA	NA	NA
IGK	NA	0.1734	NA	NA	NA	NA	NA	NA	0.1229	NA
IGL	NA	0.2289	NA	NA	NA	NA	NA	NA	NA	NA
TRA	NA	0.1575	NA	NA	NA	NA	NA	NA	NA	NA
TRB	NA	0.1790	NA	NA	0.0072	NA	NA	0.0135	NA	NA

Table 23. *Entropy Train Pearson Correlations with RiR in Scenario 2*

Entropy Train	1	2	3	4	5	6	7	8	9	10
IGH	NA	NA	NA	NA	NA	NA	0.9999	0.9999	1	NA
IGK	1	0.9997	NA	0.9998	0.9996	NA	1	NA	NA	1
IGL	NA	NA	0.9997	NA	NA	0.9997	0.9998	0.9998	0.9999	NA
TRA	NA	0.9998	NA	NA	NA	NA	NA	NA	NA	NA
TRB	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

Table 24. *Entropy Test Pearson Correlations with RiR in Scenario 2*

Entropy Test	1	2	3	4	5	6	7	8	9	10
IGH	NA	NA	NA	NA	NA	NA	0.1788	0.0381	0.0504	NA
IGK	0.0132	0.1746	NA	0.0847	0.0644	NA	0.1750	NA	NA	0.1019
IGL	NA	NA	0.1000	NA	NA	0.0027	0.2968	0.0397	0.0567	NA
TRA	NA	0.0435	NA	NA	NA	NA	NA	NA	NA	NA
TRB	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

Table 25. *Expression Train Pearson Correlations with RF in Scenario 2*

Expression Train	1	2	3	4	5	6	7	8	9	10
IGH	0.9975	0.9980	0.9980	0.9980	0.9984	0.9976	0.9974	0.9970	0.9987	0.9979
IGK	0.9968	0.9978	0.9981	0.9981	0.9982	0.9982	0.9971	0.9970	0.9974	0.9976
IGL	0.9974	0.9974	0.9984	0.9977	0.9975	0.9976	0.9974	0.9965	0.9963	0.9976
TRA	0.9968	0.9975	0.9978	0.9975	0.9965	0.9977	0.9984	0.9977	0.9982	0.9975
TRB	0.9967	0.9971	0.9981	0.9971	0.9962	0.9980	0.9971	0.9982	0.9982	0.9965

Table 26. *Expression Test Pearson Correlations with RF in Scenario 2*

Expression Test	1	2	3	4	5	6	7	8	9	10
IGH	0.3086	0.4553	0.1749	0.1960	0.0273	0.0386	0.0883	0.2014	0.3062	0.0330
IGK	0.0949	0.2069	0.1992	0.1884	0.1957	0.2480	0.0732	0.1464	0.0785	0.0232
IGL	0.0454	0.0236	0.0776	0.0165	0.3716	0.1564	0.2269	0.3208	0.1199	0.3512
TRA	0.1191	0.1109	0.0010	0.1912	0.0651	0.0474	0.0473	0.1011	0.0457	0.1981
TRB	0.0692	0.0507	0.1292	0.1191	0.1905	0.3571	0.0795	0.1385	0.2570	0.0176

Table 27. *Entropy Train Pearson Correlations with RF in Scenario 2*

Entropy Train	1	2	3	4	5	6	7	8	9	10
IGH	0.9976	0.9985	0.9987	0.9982	0.9967	0.9971	0.9980	0.9982	0.9981	0.9984
IGK	0.9963	0.9956	0.9975	0.9949	0.9948	0.9937	0.9948	0.9960	0.9964	0.9965
IGL	0.9956	0.9975	0.9959	0.9975	0.9957	0.9985	0.9960	0.9955	0.9979	0.9978
TRA	0.9975	0.9975	0.9978	0.9979	0.9980	0.9983	0.9970	0.9974	0.9979	0.9984
TRB	0.9976	0.9980	0.9981	0.9979	0.9986	0.9977	0.9973	0.9974	0.9949	0.9972

Table 28. *Entropy Test Pearson Correlations with RF in Scenario 2*

Entropy Test	1	2	3	4	5	6	7	8	9	10
IGH	0.1385	0.0366	0.2146	0.2982	0.0947	0.0199	0.2269	0.0133	0.1275	0.0911
IGK	0.2408	0.0404	0.1303	0.1620	0.1962	0.0431	0.1503	0.1618	0.2729	0.1065
IGL	0.1097	0.2259	0.1152	0.0695	0.0796	0.0256	0.1088	0.2031	0.3031	0.1462
TRA	0.1959	0.0788	0.1738	0.0250	0.1710	0.0962	0.0140	0.2295	0.1489	0.0376
TRB	0.0731	0.2660	0.0360	0.2505	0.0347	0.2217	0.1659	0.0697	0.2565	0.3455

Table 29. *Expression Train Pearson Correlations with NN in Scenario 2*

Expression Train	1	2	3	4	5	6	7	8	9	10
IGH	0.9831	0.9651	0.9927	0.9920	0.9862	0.9831	0.9924	0.9838	0.9836	0.9854
IGK	0.9848	0.9826	0.9876	0.9834	0.9863	0.9803	0.9871	0.9815	0.9836	0.9812
IGL	0.9661	0.9865	0.9814	0.9812	0.9788	0.9919	0.9812	0.9659	0.9885	0.9903
TRA	0.9466	0.9676	0.9562	0.9460	0.9783	0.9595	0.9816	0.9658	0.8930	0.9629
TRB	0.9701	0.9654	0.9245	0.9716	0.9660	0.8783	0.9583	0.9209	0.9304	0.9879

Table 30. *Expression Test Pearson Correlations with NN in Scenario 2*

Expression Test	1	2	3	4	5	6	7	8	9	10
IGH	0.0490	0.4735	0.1244	0.2997	0.3625	0.2651	0.1192	0.2929	0.0414	0.0361
IGK	0.1471	0.0719	0.0827	0.0469	0.0632	0.0692	0.2851	0.2098	0.2143	0.0486
IGL	0.2892	0.0845	0.2195	0.0952	0.0299	0.2358	0.1483	0.1558	0.3020	0.3154
TRA	0.1214	0.1797	0.1089	0.0804	0.1899	0.0071	0.4124	0.3201	0.1089	0.0939
TRB	0.1161	0.1011	0.0009	0.0312	0.1539	0.2832	0.2642	0.1865	0.1490	0.0319

Table 31. *Entropy Train Pearson Correlations with NN in Scenario 2*

Entropy Train	1	2	3	4	5	6	7	8	9	10
IGH	0.9933	0.9908	0.9939	0.9877	0.9842	0.9824	0.9786	0.9902	0.9925	0.9905
IGK	0.9570	0.9592	0.9621	0.9355	0.9453	0.9787	0.9470	0.9438	0.9802	0.9634
IGL	0.9856	0.9648	0.9859	0.9681	0.9734	0.9793	0.9852	0.9787	0.9770	0.9726
TRA	0.9875	0.9878	0.9926	0.9859	0.9865	0.9920	0.9894	0.9825	0.9838	0.9822
TRB	0.9816	0.9626	0.9923	0.9892	0.9833	0.9937	0.9933	0.9836	0.9907	0.9882

Table 32. *Entropy Test Pearson Correlations with NN in Scenario 2*

Entropy Test	1	2	3	4	5	6	7	8	9	10
IGH	0.2424	0.1805	0.0458	0.2363	0.1504	0.0524	0.0983	0.0586	0.0527	0.1081
IGK	0.2819	0.0787	0.1230	0.0214	0.2519	0.3619	0.0933	0.0382	0.1620	0.0685
IGL	0.0640	0.4381	0.0448	0.1776	0.1493	0.0112	0.1116	0.0381	0.2572	0.1729
TRA	0.0026	0.1334	0.0221	0.1371	0.0231	0.2530	0.0117	0.0919	0.0798	0.0161
TRB	0.0253	0.0017	0.4040	0.1765	0.1072	0.0733	0.1733	0.1172	0.1884	0.0138

2.3 SCENARIO 3

Table 33. *Expression Train Pearson Correlations with ENET in Scenario 3*

Expression Train	1	2	3	4	5	6	7	8	9	10
IGH	0.9999	NA	NA	1	NA	1	0.9990	0.6540	NA	0.9508
IGK	NA	1	NA	NA	NA	1	0.6369	0.7656	0.9279	NA
IGL	NA	0.6372	NA	NA	0.6083	1	NA	NA	NA	0.8049
TRA	0.6223	1	1	0.9999	0.8833	NA	NA	NA	NA	1
TRB	0.9987	0.9946	1	0.9941	NA	1	NA	NA	0.5849	0.5963

Table 34. *Expression Test Pearson Correlations with ENET in Scenario 3*

Expression Test	1	2	3	4	5	6	7	8	9	10
IGH	0.2260	NA	NA	0.0253	NA	0.1809	0.1696	0.1629	NA	0.0699
IGK	NA	0.0024	NA	NA	NA	0.0295	0.0522	0.0161	0.0398	NA
IGL	NA	0.1427	NA	NA	0.0119	0.3309	NA	NA	NA	0.2451
TRA	0.0682	0.0593	0.1149	0.3119	0.0567	NA	NA	NA	NA	0.2794
TRB	0.2488	0.2704	0.0150	0.1746	NA	0.2929	NA	NA	0.0565	0.1416

Table 35. *Entropy Train Pearson Correlations with ENET in Scenario 3*

Entropy Train	1	2	3	4	5	6	7	8	9	10
IGH	NA	0.9992	0.8039	0.9903	0.9966	NA	NA	NA	0.9884	0.9751
IGK	0.9925	0.9929	1	0.9754	NA	NA	0.9534	0.7156	NA	NA
IGL	0.9877	NA	0.9768	1	1	NA	0.9202	1	1	0.7686
TRA	NA	NA	0.6228	NA	1	0.8391	0.9997	NA	1	NA
TRB	0.9062	NA	0.9402	0.9001	NA	NA	NA	1	NA	NA

Table 36. *Entropy Test Pearson Correlations with ENET in Scenario 3*

Entropy Test	1	2	3	4	5	6	7	8	9	10
IGH	NA	0.2203	0.3288	0.0006	0.5332	NA	NA	NA	0.0642	0.3099
IGK	0.2582	0.3019	0.0972	0.1261	NA	NA	0.1556	0.0677	NA	NA
IGL	0.3175	NA	0.1752	0.0675	0.1654	NA	0.0646	0.0743	0.0699	0.3353
TRA	NA	NA	0.1251	NA	0.2046	0.4268	0.0669	NA	0.3231	NA
TRB	0.0589	NA	0.1528	0.0269	NA	NA	NA	0.1204	NA	NA

Table 37. *Expression Train Pearson Correlations with RiR in Scenario 3*

Expression Train	1	2	3	4	5	6	7	8	9	10
IGH	NA	0.9998	NA	0.9998	0.9999	NA	NA	NA	NA	1
IGK	NA	NA	NA	1	NA	0.9996	NA	NA	NA	NA
IGL	0.9998	NA	1	1	NA	1	0.9998	1	1	1
TRA	NA	1	1	NA	0.9998	NA	NA	NA	NA	NA
TRB	0.9998	NA	0.9998	NA	0.9998	NA	NA	NA	0.9998	1

Table 38. *Expression Test Pearson Correlations with RiR in Scenario 3*

Expression Test	1	2	3	4	5	6	7	8	9	10
IGH	NA	0.9839	NA	0.9901	0.9851	NA	NA	NA	NA	0.9867
IGK	NA	NA	NA	0.9849	NA	0.9864	NA	NA	NA	NA
IGL	0.9884	NA	0.9910	0.9885	NA	0.9906	0.9864	0.9883	0.9903	0.9932
TRA	NA	0.9951	0.9860	NA	0.9928	NA	NA	NA	NA	NA
TRB	0.9834	NA	0.9780	NA	0.9810	NA	NA	NA	0.9880	0.9898

Table 39. *Entropy Train Pearson Correlations with RiR in Scenario 3*

Entropy Train	1	2	3	4	5	6	7	8	9	10
IGH	NA	NA	NA	NA	NA	NA	1	NA	NA	NA
IGK	0.9997	NA	10.000	10.000	10.000	10.000	0.9999	0.9999	0.9998	0.9999
IGL	0.9999	1	1	0.9998	0.9999	1	0.9998	0.9998	0.9999	1
TRA	NA	1	NA	NA	0.9998	0.9997	NA	NA	0.9998	NA
TRB	1	1	1	0.9999	1	NA	NA	NA	NA	NA

Table 40. *Entropy Test Pearson Correlations with RiR in Scenario 3*

Entropy Test	1	2	3	4	5	6	7	8	9	10
IGH	NA	NA	NA	NA	NA	NA	0.9953	NA	NA	NA
IGK	0.9793	NA	0.9857	0.9858	0.9836	0.9940	0.9877	0.9730	0.9881	0.9705
IGL	0.9928	0.9922	0.9902	0.9903	0.9948	0.9905	0.9936	0.9863	0.9866	0.9907
TRA	NA	0.9874	NA	NA	0.9874	0.9888	NA	NA	0.9854	NA
TRB	0.9953	0.9956	0.9960	0.9928	0.9941	NA	NA	NA	NA	NA

Table 41. *Expression Train Pearson Correlations with RF in Scenario 3*

Expression Train	1	2	3	4	5	6	7	8	9	10
IGH	0.9972	0.9978	0.9974	0.9982	0.9981	0.9980	0.9979	0.9986	0.9985	0.9983
IGK	0.9977	0.9974	0.9975	0.9979	0.9972	0.9981	0.9984	0.9974	0.9975	0.9974
IGL	0.9959	0.9981	0.9972	0.9974	0.9975	0.9978	0.9981	0.9978	0.9978	0.9971
TRA	0.9968	0.9969	0.9973	0.9978	0.9958	0.9966	0.9977	0.9970	0.9981	0.9975
TRB	0.9978	0.9978	0.9967	0.9957	0.9970	0.9980	0.9970	0.9964	0.9968	0.9972

Table 42. *Expression Test Pearson Correlations with RF in Scenario 3*

Expression Test	1	2	3	4	5	6	7	8	9	10
IGH	0.5571	0.7102	0.5986	0.1761	0.6622	0.7159	0.3810	0.6847	0.5372	0.5160
IGK	0.1365	0.2643	0.2788	0.5038	0.5034	0.4422	0.5688	0.5736	0.3114	0.5923
IGL	0.2715	0.4012	0.4048	0.3324	0.6300	0.2908	0.4549	0.5031	0.4994	0.5043
TRA	0.4685	0.5566	0.3745	0.4821	0.1692	0.2197	0.5856	0.6385	0.6036	0.4572
TRB	0.3669	0.5096	0.5355	0.2030	0.4392	0.5930	0.1150	0.5485	0.5540	0.5948

Table 43. *Entropy Train Pearson Correlations with RF in Scenario 3*

Entropy Train	1	2	3	4	5	6	7	8	9	10
IGH	0.9970	0.9962	0.9983	0.9984	0.9978	0.9973	0.9975	0.9972	0.9967	0.9983
IGK	0.9966	0.9951	0.9961	0.9967	0.9961	0.9963	0.9963	0.9954	0.9958	0.9970
IGL	0.9974	0.9941	0.9985	0.9958	0.9970	0.9976	0.9968	0.9946	0.9978	0.9968
TRA	0.9982	0.9973	0.9973	0.9969	0.9976	0.9981	0.9977	0.9979	0.9977	0.9978
TRB	0.9982	0.9970	0.9975	0.9981	0.9980	0.9972	0.9974	0.9975	0.9951	0.9981

Table 44. *Entropy Test Pearson Correlations with RF in Scenario 3*

Entropy Test	1	2	3	4	5	6	7	8	9	10
IGH	0.7341	0.6214	0.7249	0.2689	0.5722	0.4168	0.4719	0.5146	0.4062	0.7676
IGK	0.2040	0.4099	0.5078	0.4308	0.3061	0.4293	0.2357	0.4023	0.1677	0.6926
IGL	0.5167	0.1256	0.7077	0.1484	0.4099	0.1615	0.4979	0.5672	0.3954	0.3576
TRA	0.4546	0.6628	0.6729	0.6897	0.5467	0.5591	0.6216	0.2260	0.5096	0.4722
TRB	0.2062	0.4156	0.2594	0.3533	0.2486	0.3545	0.6464	0.2246	0.5244	0.7081

Table 45. *Expression Train Pearson Correlations with NN in Scenario 3*

Expression Train	1	2	3	4	5	6	7	8	9	10
IGH	0.9972	0.9978	0.9974	0.9982	0.9981	0.9980	0.9979	0.9986	0.9985	0.9983
IGK	0.9977	0.9974	0.9975	0.9979	0.9972	0.9981	0.9984	0.9974	0.9975	0.9974
IGL	0.9959	0.9981	0.9972	0.9974	0.9975	0.9978	0.9981	0.9978	0.9978	0.9971
TRA	0.9968	0.9969	0.9973	0.9978	0.9958	0.9966	0.9977	0.9970	0.9981	0.9975
TRB	0.9978	0.9978	0.9967	0.9957	0.9970	0.9980	0.9970	0.9964	0.9968	0.9972

Table 46. *Expression Test Pearson Correlations with NN in Scenario 3*

Expression Test	1	2	3	4	5	6	7	8	9	10
IGH	0.5571	0.7102	0.5986	0.1761	0.6622	0.7159	0.3810	0.6847	0.5372	0.5160
IGK	0.1365	0.2643	0.2788	0.5038	0.5034	0.4422	0.5688	0.5736	0.3114	0.5923
IGL	0.2715	0.4012	0.4048	0.3324	0.6300	0.2908	0.4549	0.5031	0.4994	0.5043
TRA	0.4685	0.5566	0.3745	0.4821	0.1692	0.2197	0.5856	0.6385	0.6036	0.4572
TRB	0.3669	0.5096	0.5355	0.2030	0.4392	0.5930	0.1150	0.5485	0.5540	0.5948

Table 47. *Entropy Train Pearson Correlations with NN in Scenario 3*

Entropy Train	1	2	3	4	5	6	7	8	9	10
IGH	0.9970	0.9962	0.9983	0.9984	0.9978	0.9973	0.9975	0.9972	0.9967	0.9983
IGK	0.9966	0.9951	0.9961	0.9967	0.9961	0.9963	0.9963	0.9954	0.9958	0.9970
IGL	0.9974	0.9941	0.9985	0.9958	0.9970	0.9976	0.9968	0.9946	0.9978	0.9968
TRA	0.9982	0.9973	0.9973	0.9969	0.9976	0.9981	0.9977	0.9979	0.9977	0.9978
TRB	0.9982	0.9970	0.9975	0.9981	0.9980	0.9972	0.9974	0.9975	0.9951	0.9981

Table 48. *Entropy Test Pearson Correlations with NN in Scenario 3*

Entropy Test	1	2	3	4	5	6	7	8	9	10
IGH	0.7341	0.6214	0.7249	0.2689	0.5722	0.4168	0.4719	0.5146	0.4062	0.7676
IGK	0.2040	0.4099	0.5078	0.4308	0.3061	0.4293	0.2357	0.4023	0.1677	0.6926
IGL	0.5167	0.1256	0.7077	0.1484	0.4099	0.1615	0.4979	0.5672	0.3954	0.3576
TRA	0.4546	0.6628	0.6729	0.6897	0.5467	0.5591	0.6216	0.2260	0.5096	0.4722
TRB	0.2062	0.4156	0.2594	0.3533	0.2486	0.3545	0.6464	0.2246	0.5244	0.7081

3. STANDARD DEVIATION VALUES

3.1 SCENARIO 1

Table 49. *SD in Expression variables with ENET in scenario 1*

Expression	IGH	IGK	IGL	TRA	TRB
Train	0.0000	0.0213	0.0000	0.0252	0.1384
Test	0.0684	0.1057	0.1071	0.0889	0.0949

Table 50. *SD in Entropy variables with ENET in scenario 1*

Entropy	IGH	IGK	IGL	TRA	TRB
Train	0.0002	0.0000	0.0000	0.0325	0.0254
Test	0.1342	0.0795	0.1134	0.1222	0.1124

Table 51. *SD in Expression variables with RiR in scenario 1*

Expression	IGH	IGK	IGL	TRA	TRB
Train	0.00017	0.00019	0.00018	0.00010	0.00017
Test	0.0022	0.0027	0.0020	0.0031	0.0027

Table 52. *SD in Entropy variables with RiR in scenario 1*

Entropy	IGH	IGK	IGL	TRA	TRB
Train	Train	0.000079	0.000211	0.000095	0.000176
Test	Test	0.0008	0.0044	0.0015	0.0011

Table 53. *SD in Expression variables with RF in scenario 1*

Expression	IGH	IGK	IGL	TRA	TRB
Train	0.0002	0.0004	0.0006	0.0004	0.0006
Test	0.0634	0.0578	0.0736	0.0713	0.0525

Table 54. *SD in Entropy variables with RF in scenario 1*

Entropy	IGH	IGK	IGL	TRA	TRB
Train	0.0007	0.0007	0.001	0.0005	0.0003
Test	0.0765	0.1654	0.2359	0.0890	0.0516

Table 55. *SD in Expression variables with NN in scenario 1*

Expression	IGH	IGK	IGL	TRA	TRB
Train	0.0052	0.0028	0.0060	0.0050	0.0088
Test	0.0231	0.0177	0.0228	0.0328	0.0591

Table 56. *SD in Expression variables with NN in scenario 1*

Entropy	IGH	IGK	IGL	TRA	TRB
Train	0.0029	0.0064	0.0052	0.0023	0.0021
Test	0.0160	0.0505	0.0208	0.0220	0.0155

3.2 SCENARIO 2

Table 57. *SD in Expression variables with ENET in scenario 2*

Expression	IGH	IGK	IGL	TRA	TRB
Train	0.1068	0.0038	0.2007	0.1294	0.2058
Test	0.0881	0.0622	0.085	0.0671	0.0892

Table 58. *SD in Entropy variables with ENET in scenario 2*

Entropy	IGH	IGK	IGL	TRA	TRB
Train	0.1635	0.156	0.0568	0.3183	0.205
Test	0.059	0.0525	0.0707	0.0789	0.099

Table 59. *SD in Expression variables with RiR in scenario 2*

Expression	IGH	IGK	IGL	TRA	TRB
Train	NA	0.0001	NA	NA	0.0002
Test	NA	0.0357	NA	NA	0.0974

Table 60. *SD in Entropy variables with RiR in scenario 2*

Entropy	IGH	IGK	IGL	TRA	TRB
Train	0,0001	0,0002	0,0001	NA	NA
Test	0.0779	0.0636	0.1159	NA	NA

Table 61. *SD in Expression variables with RF in scenario 2*

Expression	IGH	IGK	IGL	TRA	TRB
Train	0.0005	0.0005	0.0006	0.0006	0.0008
Test	0.1423	0.0736	0.1379	0.0645	0.103

Table 62. *SD in Entropy variables with RF in scenario 2*

Entropy	IGH	IGK	IGL	TRA	TRB
Train	0.0006	0.0011	0.0011	0.0004	0.001
Test	0.0954	0.0756	0.0831	0.077	0.1117

Table 63. *SD in Expression variables with NN in scenario 2*

Expression	IGH	IGK	IGL	TRA	TRB
Train	0.008	0.0025	0.0091	0.0249	0.033
Test	0.153	0.0848	0.0998	0.1208	0.0958

Table 64. *SD in Entropy variables with NN in scenario 2*

Entropy	IGH	IGK	IGL	TRA	TRB
Train	0.0051	0.0147	0.0074	0.0036	0.0092
Test	0.076	0.1143	0.128	0.0799	0.1189

3.3 SCENARIO 3

Table 65. *SD in Expression variables with ENET in scenario 3*

Expression	IGH	IGK	IGL	TRA	TRB
Train	0.1385	0.1599	0.1804	0.1520	0.1986
Test	0.0755	0.0195	0.1374	0.1164	0.1073

Table 66. *SD in Entropy variables with ENET in scenario 3*

Entropy	IGH	IGK	IGL	TRA	TRB
Train	0.0764	0.1104	0.0807	0.1660	0.0458
Test	0.1935	0.0928	0.1127	0.1463	0.0572

Table 67. *SD in Expression variables with RiR in scenario 3*

Expression	IGH	IGK	IGL	TRA	TRB
Train	0.0001	0.0003	0.0001	0.0001	0.0001
Test	0.0027	0.0011	0.0021	0.0047	0.0049

Table 68. *SD in Entropy variables with RiR in scenario 3*

Entropy	IGH	IGK	IGL	TRA	TRB
Train	NA	0.0001	0.0001	0.0001	0.0000
Test	NA	0.0075	0.0028	0.0014	0.0013

Table 69. *SD in Expression variables with RF in scenario 3*

Expression	IGH	IGK	IGL	TRA	TRB
Train	0,0004	0,0004	0,0006	0,0007	0,0007
Test	0.1686	0.1585	0.1114	0.1590	0.1676

Table 70. *SD in Entropy variables with RF in scenario 3*

Entropy	IGH	IGK	IGL	TRA	TRB
Train	0.0007	0.0006	0.0014	0.0004	0.0009
Test	0.1641	0.1573	0.1949	0.1389	0.1783

Table 71. *SD in Expression variables with NN in scenario 3*

Expression	IGH	IGK	IGL	TRA	TRB
Train	0.0144	0.0100	0.0105	0.0230	0.0228
Test	0.0294	0.0364	0.0298	0.0601	0.0490

Table 72. *SD in Entropy variables with NN in scenario 3*

Entropy	IGH	IGK	IGL	TRA	TRB
Train	0.0069	0.0192	0.0038	0.0063	0.0069
Test	0.0196	0.1011	0.0285	0.0183	0.0257