mathematics

MDPI

Article

GASVeM: A New Machine Learning Methodology for Multi-SNP Analysis of GWAS Data Based on Genetic Algorithms and Support Vector Machines

Fidel Díez Díaz [1], Fernando Sánchez Lasheras [2,*], Víctor Moreno [3], Ferran Moratalla-Navarro [3], Antonio José Molina de la Torre [4] and Vicente Martín Sánchez [5]

1 CTIC Technological Centre, W3C Spain Office Host, Ada Byron 39, 33203 Gijón, Spain; fidel.diez@fundacionctic.org
2 Department of Mathematics, Faculty of Sciences, Universidad de Oviedo, 33007 Oviedo, Spain
3 Oncology Data Analytics Program, Catalan Institute of Oncology (ICO), Colorectal Cancer Group, ONCOBELL Program, Bellvitge Biomedical Research Institute (IDIBELL), Consortium for Biomedical Research in Epidemiology and Public Health (CIBERESP), Department of Clinical Sciences, Faculty of Medicine, University of Barcelona, 08908 Barcelona, Spain; v.moreno@iconcologia.net (V.M.); fmoratalla@iconcologia.net (F.M.-N.)
4 IBIOMED, University of Leon, Vegazana Campus, 24400 León, Spain; ajmolt@unileon.es
5 CIBERESP, University of Leon, Vegazana Campus, 24400 León, Spain; vicente.martin@unileon.es
* Correspondence: sanchezfernando@uniovi.es; Tel.: +34-98-510-3338

Abstract: Genome-wide association studies (GWAS) are observational studies of a large set of genetic variants in an individual's sample in order to find if any of these variants are linked to a particular trait. In the last two decades, GWAS have contributed to several new discoveries in the field of genetics. This research presents a novel methodology to which GWAS can be applied to. It is mainly based on two machine learning methodologies, genetic algorithms and support vector machines. The database employed for the study consisted of information about 370,750 single-nucleotide polymorphisms belonging to 1076 cases of colorectal cancer and 973 controls. Ten pathways with different degrees of relationship with the trait under study were tested. The results obtained showed how the proposed methodology is able to detect relevant pathways for a certain trait: in this case, colorectal cancer.

Keywords: machine learning; support vector machines; genetic algorithms; genome-wide association studies; single nucleotide polymorphism; pathways analysis

Citation: Díez Díaz, F.; Sánchez Lasheras, F.; Moreno, V.; Moratalla-Navarro, F.; Molina de la Torre, A.J.; Martín Sánchez, V. GASVeM: A New Machine Learning Methodology for Multi-SNP Analysis of GWAS Data Based on Genetic Algorithms and Support Vector Machines. *Mathematics* **2021**, *9*, 654. https://doi.org/10.3390/math9060654

Academic Editor: Seungmin Rho

Received: 17 January 2021
Accepted: 12 March 2021
Published: 18 March 2021

1. Introduction

The results of the Human Genome Project [1] and the International HapMap Project [2] made it possible to find genes linked to traits and health problems. Genome-wide association studies (GWAS) have contributed to several new discoveries in human genetics. GWAS exploit the fact that genetic variants that are close together tend to be statistically correlated, something which in genetics is known as linkage disequilibrium [3]. The advances in genome arrays of genetic variations have led to the discovery of many DNA variants associated with complex traits such as those related to diseases.

Nowadays, one of the main criticisms of GWAS is that to date, most of the discoveries have not been applied in a clinical practice [4], but despite this obvious drawback, GWAS do have a great relevance. As an example, it can be said that until the development of GWAS it was not possible to find any gene linked to schizophrenia [5].

Not only are GWAS of interest for the discovery of robust associations, but they also give information about the nature of variations in traits and have contributed to the discovery of new biological knowledge about how DNA variations can affect gene regulation. The variations in the human genome are mainly down to two causes: point

mutations and structural variation [6]. When a point mutation occurs, a DNA base is replaced by another. In this case of structural variations like this, changes are wider and can range from small insertions or deletions to large chromosomal rearrangements [6]. Each kind of structural variation has different rates of mutation and evolution and their role in phenotypic variation is well-known.

The first GWAS were published in 2005 and 2006 [7,8]. Both were able to find common variants associated to age-related macular degeneration. GWAS can go beyond the candidate gene studies. The reasons for the greatest capabilities of GWAS are twofold. On the one hand, while in candidate gene studies only a few selected single nucleotide polymorphism (SNPs) are considered, a GWAS means simultaneously studying a large number of SNPs representative of whole-genome genetic variation; it is also worth remarking that GWAS are considered to be "hypothesis free". In other words, they are able to look for common risk effects looking at SNPs located across all or a considerable part of the genome without any list of a priori loci [9].

GWAS can survey the role of common genetic variations in complex human diseases. It was expected that GWAS would have the advantage of not relying on prior knowledge of biological pathways compared with "candidate genes" studies [10]. This advantage allows GWAS to overcome the bias of "candidate genes" studies. Biological pathways can be defined as a group of genes that are related from a functional point of view.

The major challenge of GWAS data analysis is the polygenic architecture of complex diseases. This means that in the presence of numerous variants with small or moderate effects, a large sample size is needed for both association mapping and risk prediction. However, sample recruitment can be expensive and time-consuming.

The key step for the validation of the association between genetic variants and complex human diseases is the replication of findings in independent samples. Replication of newly reported associations is usually considered to be the most reliable validation of GWAS discoveries.

The first GWAS considered a link between SNPs and phenotype, but one SNP at a time [11]. Nowadays studies make use of more complex analysis that includes multivariate analysis [12] or machine learning approaches [13,14]. GWAS have allowed for a better understanding of the genetic components of many complex traits.

The aim of this research is to explore a new methodology based on machine learning that is able to find sets of SNPs selected from pathways that can differentiate cases from controls. This method is based on genetic algorithms and support vector machines. It is called genetic algorithms support vector machines methodology (GASVeM). In classical Mendelian genetics, epistasis refers to the masking of genotypic effects at one locus by genotypes of another [15]. In quantitative genetics, epistasis can refer to a modification of the additive and/or dominance effects of the interacting loci, and the proposed methodology can deal with epistasis as understood in quantitative genetics. The performance of the new proposed methodology has been checked with the help of a GWAS database and some well-known pathways.

2. Materials and Methods

2.1. Support Vector Machines

Support vector machines (SVM) are a supervised-learning classification technique that has shown its ability in dealing with classification [16] and regression problems [17,18]. In the case of the present research, SVM is employed for binary classification in cases and controls. Let us suppose they are denoted by $\{-1, +1\}$. Therefore, predictors of the form $f : \mathbb{R}^D \to \{-1, +1\}$ are considered.

The SNPs of each member of the population are represented by a vector $x_n \in \mathbb{R}^D$ where D is the number of SNPs employed in the study and the case or control label is given by y_n. Given a training dataset consisting of pairs $\{(x_1, y_1), (x_2, y_2) \dots (x_n, y_n)\}$ with the objective being to train an SVM model with the lowest classification error.

Let $x_i \in \mathbb{R}^D$ be a data sample, and consider the function $f, \mathbb{R}^D \to \mathbb{R}$ in such a way that $x \mapsto \langle w, x \rangle + b$ are $w \in \mathbb{R}^D$ and $b \in \mathbb{R}$ the hyperplane that separates the two classes in the binary classification problem can be written as $\{x \in \mathbb{R}^D : f(x) = 0\}$.

Please note that w is a normal vector for the hyperplane and b the intercept.

When training the classifier, we want to ensure that the examples with positive labels (cases) are on the positive side of the hyperplane, while those with negative labels are on the negative side. These two conditions can be expressed as $y_n(\langle w, x_n \rangle + b) \geq 0$ where $y_n = -1$ or 1.

To find a unique solution, one idea is to choose the separating hyperplane that maximizes the margin between the cases and controls. The margin represents the distance of the hyperplane to the closest examples of cases and controls, respectively, in those cases in which it would be possible to assume that the dataset is linearly separable [19], but this is not the case of the present research.

Let us consider one individual x_i, which without loss of generality can be considered as a case and labeled as $+1$. This case observation should be on the positive side of the hyperplane $\langle w, x_i \rangle + b > 0$. The distance of x_i to the hyperplane is given by the distance of the orthogonal projection of the point to the plane. Using vector addition, x_i can be expressed as follows:

$$x_i = \alpha'_i + r\frac{\omega}{\|\omega\|}$$

In other words, all the case observations must be at least a distance r from the hyperplane. It can be expressed by the following equation:

$$y_n(\langle w, x_n \rangle + b) \geq r$$

And the optimization problem to be solved can be expressed as:

$$\max r$$

Subject to: $y_n(\langle w, x_n \rangle + b) \geq r$

$$\|\omega\| = 1,\ r > 0$$

Can then be interpreted as the maximization of r while ensuring that the case observations lie on the correct side of the hyperplane. This is usually called the margin maximization parameter.

In these cases, like in the one in the present problem where data are not linearly separable, some examples would fall into the margin region or even on the wrong side of the hyperplane. This model is called soft margin SVM and makes use of the following equations:

$$\min \frac{1}{2}\|\omega\|^2 + C\sum_{n=1}^{N} \xi_{n^k}$$

Subject to: $y_n(\langle w, x_n \rangle + b) = 1 - \xi_n, \xi_n \geq 0$.

ξ_n is the slack variable corresponding to each observation. The parameter $C > 0$ trades off the size of the margin and the total amount of slack that we have, and is called the regularization parameter.

The margin term $\|\omega\|^2$ is called the regularizer.

The convex duality via Lagrange Multipliers of the previous formula can be expressed as follows:

$$\alpha(\omega, b, \xi, \alpha, \gamma) = \frac{1}{2}\|\omega\|^2 + C\sum_{n=1}^{N} \xi_n - \sum_{n=1}^{N} \alpha_n(y_n(\langle w, x_n \rangle + b) - 1 + \xi_n) - \sum_{n=1}^{N} \gamma_n \xi_n$$

When the Lagrangian is differentiated with respect to the three primal variables, ω, b and ξ, the following result is obtained:

$$\frac{\partial \alpha}{\partial \omega} = \omega^T - \sum_{n=1}^{N} \alpha_n y_n x_n^T \frac{\partial \alpha}{\partial b} = \sum_{n=1}^{N} \alpha_n y_n \frac{\partial \alpha}{\partial \xi_n} = C - \alpha_n - \gamma_n$$

Setting the three partial derivatives to zero, it would be possible to express the previous equations by means of their dual negative equations. This converts the problem from a maximization into a minimization one:

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} y_i y_j \alpha_i \alpha_j \langle x_i, x_j \rangle - \sum_{i=1}^{N} \alpha_i$$

Subject to: $\sum_{i=1}^{N} y_i \alpha_i = 0$.

$0 \leq \alpha_i \leq C$ for all $i = \{, \ldots, N\}$.

Although SVM requires linearity for classification, the use of kernels makes this methodology also useful for non-linear problems. A kernel can be defined as a function $K : X \times X \to \mathbb{R}$ for which there is a Hilbert space H and $\phi : X \to H$ is feature map such that:

$$k\langle x_i, x_j \rangle = \left\langle \phi(x_i), \phi\left(x_j\right) \right\rangle_H$$

Kernels must be symmetric and positive semidefinite functions in such a way that the kernel matrix K is symmetric and positive semidefinite. The use of different kernels has different effects on the separating hyperplanes. In this research lineal, polynomial, radial basis functions and sigmoid kernels have all been tested. These kernel functions can be expressed as follows [19]:

Lineal: $k\left(x_i : x_j\right) = x_i^T x_j$

Polynomial: $k\left(x_i, x_j\right) = \left(\gamma x_i^T x_j + C\right)^P$

Radial basis function: $k\left(x_i, x_j\right) = \exp\left[-\gamma \|x_i - x_j\|^2\right]$

Sigmoid: $\tan b\left(\gamma x_i^T x_j + k\right)$

2.2. Genetic Algorithms

Genetic algorithms can be defined as biologically inspired methods for optimization [20]. The foundations of genetic algorithms can be found in the works of Holland [21], Rechenberg [22] and Schwefel [23].

For their initialization, genetic algorithms require an initial set of candidate solutions for the optimization problem to be solved. Table 1 shows the pseudocode of a genetic algorithm. As can be observed in the table, the first step involves creating an initial population. Data representation and how the initial population is created both have a great importance on the genetic algorithm performance. The second operation performed is the crossover.

Table 1. Pseudocode of a genetic algorithm.

1.	initialize population
2.	repeat
3.	repeat
4.	crossover
5.	mutation
6.	phenotype mapping
7.	fitness computation
8.	until population complete
9.	selection of parental population
10.	until termination condition

A non-deterministic crossover function can be defined as $X : \Omega \times \Omega \to \Omega$. The result of $X(x_i, x_j)$ gives a new population member with the same length as x_i and x_j and as such, all their elements belong either to x_i or x_j with a certain probability.

Mutation is employed to inject new strings into the next generation [24], which gives the genetic algorithm the ability to search beyond the confines of the initial population. The mutation function can be expressed as: $\mu : \Omega \to \Omega$. It is like a crossover: a non-deterministic function that assigns to each string member a certain probability of being randomly changed. The fitness computation assigns a value to each member of the population that represents how well they fit to the problem to be solved. Those individuals with the most favorable results of the fitness function are more likely to be selected as parents of the next-generation offspring.

This process is repeated until any termination condition is reached. In the case of the present research, there are certain key parameters to be controlled in order to obtain a fine-tuned version of the algorithm. These parameters are the number of iterations, the population size, crossover and probability of mutation.

2.3. The Proposed Algorithm

The algorithm proposed in the present research makes use of both genetic algorithms and support vector machines in order to find out whether a certain pathway, which in this context can be considered in the same way as a set of SNPs, is able to identify cases and controls for a certain trait or illness.

Figure 1 shows the flowchart of the proposed algorithm. The first step involves selecting the subset of SNPs that belongs to the pathway under analysis. This means that from the total number of SNPs included in the database, the information required for the analysis is reduced to a selected subset of SNPs of all the members of the population. In other words, the SNPs chosen are only those that belong to the pathway to be studied.

The members of the genetic algorithm (GA) population for this analysis are strings of "1s" and "0s" that indicate which SNPs will form a part of the SVM model to be computed. Please note that "1" means that the SNP will take part of the SVM model and "0" that it will not. In the case of the present research, each member of the GA population has the same length as the number of SNPs that constitute the pathway under analysis. Please note that each GA population has several members, and an SVM model is trained for each one.

All the classification SVM models are trained using as input variables the SNPs with the "1" value and as output, the variable trait that indicates which elements are cases and which are controls. As may be seen in the flowchart, the initial population is formed by rows from an identity matrix selected in a random way up to the completion of the total number of individuals required for the GA population. This means that in the initial population, only one SNP is active in each population member. In other words, it means that after selecting as input information only those SNPs that belong to the pathway under analysis, the initial population is formed by individuals in which only one of those SNPs is active and, afterwards, the different SNPs that belong to the subset that is being employed are switched on and off with the aim of improving the results of the fitness function. The reason for choosing only one SNP in each member of the initial population is that the goal is to get the maximum values of the fitness function while making use of the minimum number of SNPs required and to allow for the importance of every single SNP to be taken into account individually.

In the following populations, the number of SNPs selected in each population member can be more than one as each one of the population members evolves, taking into account genetic algorithms rules in search of the maximization of the value of the fitness function. The fitness function consists of calculating the area under the ROC curve [25] that is obtained when data are classified, making use of the SVM calculated for that member of the population (string of "0s" and "1s") using the active SNPs as independent variables and as dependent variables whether or not the individual suffers from a certain trait, which in the case of the present research is colorectal cancer. In order to avoid problems related to

epistasis, members of the population that choose more than one SNP from the same gene have a value of 0 assigned to their fitness function.

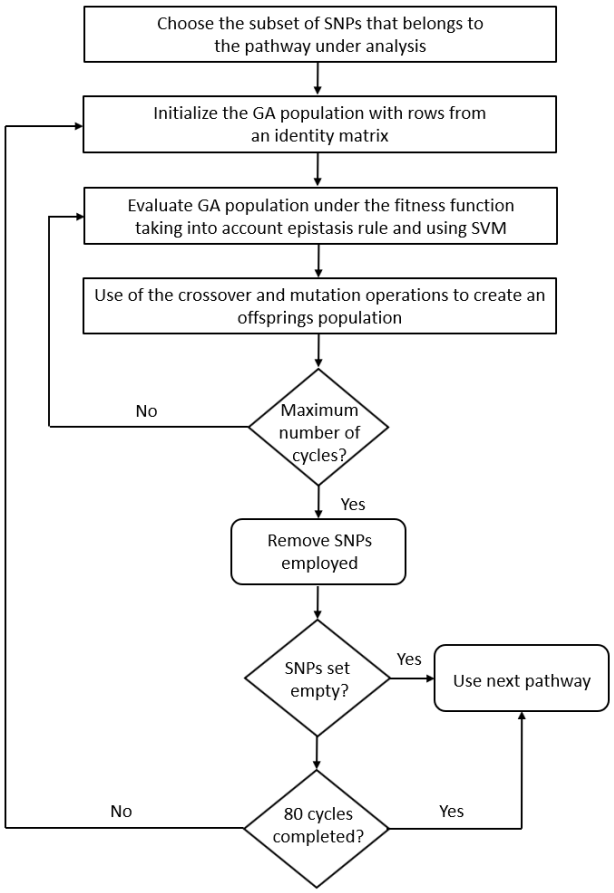

Figure 1. Flowchart of the proposed algorithm.

When the stop criterion is reached, i.e., the maximum number of cycles allowed, the area under the ROC curve (AUC)value achieved is recorded and the SNPs employed are deleted from the set of SNPs. The process starts again looking for a new SNPs subset for which the AUC value can be as high as possible. The process is then repeated until the SNP set is empty or until a total of 80 cycles have been completed. Afterwards, the same process is repeated 1000 times, making use of permutations of cases and controls labels. Please note that in this field, some of the existing literature recommends 10,000 permutations [26] while in classical literature the number of permutations considered in most of the papers for estimating the power of a permutation test was 1000 [27–29] and in some of them only 500 [30,31]. It is also worth noting that previous research [32,33] stated that 1000 is a reasonable number of permutations for a test at the 5% level of significance. Finally, in the field of genomic studies there is software that considers that permutation values from 1000 can feasibly be employed in GWAS [34].

2.4. Design of Experiments

Nowadays, it is well-known that there are no optimal parameter values valid for all problems [20]. This statement is part of the no-free-lunch theorem [35], which states that there is no overall superior optimization algorithm capable of solving every kind of optimization problem.

Parameter tuning strategies treat the parametrization of GA as an optimization problem. In the case of the present research, the GA parameters are tuned with the help of design of experiments methodology (DOE), but there are other possible approaches. For example, the parameter calibration process can be supported on statistical methods [36], a

deterministic control whereby an extended scheme is used to control the parameters [37], adaptive control strategies like Rechenberg's mutation rate [22] or self-adaption [20]. In the case of the present research, DOE was chosen.

DOE is a statistical methodology that is employed in order to find relationships among variables that affect certain processes [38]. In other words, DOE makes it possible to see how a simultaneous change in more than one variable will affect the output variable. Although DOE can be applied to both categorical and continuous variables, in the case of the present research all the variables under study with DOE are continuous.

A design of experiments was performed to select the most suitable values of the GA parameters for the algorithm proposed in the present research, using the colorectal cancer pathway as a reference. Furthermore, for designing the experiment, a full factorial design with center points was employed. In a full factorial design, all possible combinations of factor levels are used [39]. For each point, experiments were repeated three times. Please note that each individual experimental setting is referred to as a run and the response measured is called an observation. The present work made use of DOE for the fine-tuning of the GA algorithm. For the DOE analysis, the continuous variables of the research are considered, namely, the number of iterations of the algorithm, the population size and the values of mutation and crossover. The response value is the area under the ROC curve obtained with the SVM model that makes use of the variables selected. Each variable was measured at three different levels and, therefore, a 3^4 full factorial design with 81 experiments, each one repeated three times, was obtained. The variables employed and their corresponding level are shown in Table 2.

Table 2. Variable analyzed by means of design of experiments (DOE) methodology and values considered.

Variable	Low	Center Point	High
Number of iterations	4000	6000	8000
Population size	1000	5500	10,000
Crossover rate	0.1	0.55	1
Mutation rate	0.001	0.01	0.1

2.5. Datasets

The database employed in this study belongs to the Colorectal Cancer Transdisciplinary Study (CORECT) project. This was an observational multicentric multi-case control study performed from September 2008 to December 2013. For this research, the subset of information belonging to Leon University Hospital and Hospital of Bellvitge was employed. It contains 1076 cases of, and 973 controls for, colorectal cancer, for which the information from 370,570 SNPs was available.

The cases are incidental and histologically confirmed, with ages between 20 and 85. Those with communication disabilities, physically unable to participate or with a previous diagnosis of colorectal cancer were excluded. For their recruitment, the study staff contacted them at the selected hospitals.

Controls were randomly selected from the population lists assigned to family physicians in the catchment area of the hospitals where the cases were recruited, and with the same sex and age distribution ($\pm$5 years). All had been residing in the area of the hospital where the cases were recruited for at least 6 months before and did not present physical or communication impediments.

The protocol of the project was approved by the Ethics Committees of the institutions that took part in the study. The participation of the subjects in the study was voluntary, after signing an informed consent. The confidentiality of the data is guaranteed by eliminating the personal identifiers in the datasets and all the files that include information about the subjects, complying with Spain's Organic Law 15/1999. Please also note that the files have been registered with the Spain Data Protection Agency (Number 2102672171). Access to this information for other researchers is allowed on request.

For the present study, ten different pathways were selected from the KEGG database [40–42] so as to include, on the one hand, pathways for which there was already significant scientific evidence that they were associated with the trait analyzed; on another hand, pathways for which the evidence indicated that their association with the trait was improbable, and finally others for which the evidence was inconclusive. Once the ten pathways to be included had been chosen, all the SNPs that belonged to the genes considered in each one of the pathways under analysis were retrieved from the database.

3. Results

After having fixed the GA parameters to a population size of 5500, 6000 iterations for each cycle and a 100% crossover with a mutation rate of 1%, the proposed algorithm was applied to 10 different pathways.

3.1. Design of Experiments

The variables employed in this design of experiments were number of iterations, population size, crossover, and mutation rate. The values tested are presented in Table 2. Each of the combinations of variables combinations was tested three times. Figure 2 shows the main effect plots of these four variables. According to the results obtained, the number of iterations of the GA was fixed at 6000, as there is only a very small increase in the AUC value from 6000 to 8000 (about 0.1%). In the case of the population size, the value of 5500 individuals was considered to be sufficient, as increasing the number of population members to 10,000 only meant an improvement of less than 0.2% in the AUC results. For the mutation rate whose values are presented in logarithmic scale, the maximum of the three values tested was achieved for 1%, which was the center point. Finally, in the case of the crossover rate, 100% was chosen due to it giving the highest performance. For all the pathways analyzed in this research, the number of iterations is 6000, with a population size of 5500, a mutation rate of 1% and a crossover of 100%.

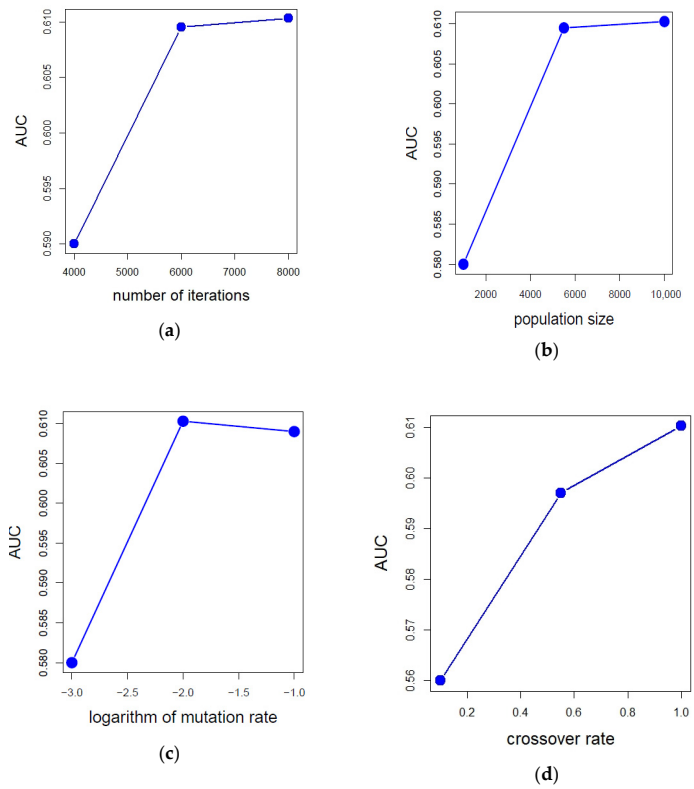

Figure 2. Main effects plots of: (**a**) number of iterations; (**b**) population size; (**c**) logarithm of mutation rate; (**d**) crossover rate.

3.2. Application of the Algorithm to Different Pathways

After having fixed the GA parameters at a population size of 5500, 6000 iterations for each cycle and a 100% crossover with a mutation rate of 1%, the proposed algorithm was applied to 10 different pathways with different degrees of relationship with the trait under study.

As can be seen in Figure 1, the first step in all the cases involves selecting as part of the initial set only those SNPs of the database taking part in the pathway under study. For example, in the case of the adipocytokine signaling pathway, it has a total of 752 SNPs and in the case of the AMP-activated protein kinase (AMPK) signaling pathway the number is 1812. The initial population of 5500 individuals is always formed by vectors in which only one element is different from zero. It indicates the SNP employed for the SVM classification in the first iteration. Starting from such an SNP, new individuals that combine zeros and ones are formed in search of those SNPs that can provide the maximum AUC value while at the same time taking into account that no more than one SNP from the same gene can be included in the same population member. This iterative process is repeated 6000 times. In the case of the AMPK signaling pathway, the AUC value obtained after this process was 0.584023, and for the adipocytokine signaling pathway it was 0.565382. When the iterations are finished, the SNPs employed in the subset with the maximum AUC obtained are removed and the process is repeated in search of the best remaining SNPs. Although this process would normally be repeated while there were still SNPs pending employment, it was halted after 80 cycles. There are two reasons for this. On the one hand, the time required for these iterations, which is quite high (34.51 s per iteration on average) and on the other hand that although we do not know in advance the exact number of SNPs involved in 80 iterations, in order to compare the results of pathways of different lengths, a number of repetitions was chosen that would be feasible for any of the pathways.

After having run the algorithm for all the pathways included in the study, the same process was repeated 1000 times for each one but permuting phenotypes of cases and controls while preserving the total number of 1076 cases of colorectal cancer and 973 controls. The results obtained were compared.

The main results are detailed in Table 3. This table presents the total number of SNPs that are included in each of the 10 pathways under analysis. As was mentioned before, the algorithm was repeated 80 times in all cases. This means that not all the SNPs were employed for the process of classification of cases and controls. For example, for the adipocytokine signaling pathway, which has a total of 752 SNPs, only 496 were employed in any of the iterations for the classification of individuals in cases and controls. Figure 3 shows the AUC values of the 80 iterations performed for the adipocytokine signaling pathway in the case of cases and controls (phenotype) and for 5 different permutations of the 1000. For the graphical representation and for the comparison of wins in pathway versus permutated, AUC values are ordered from highest to lowest. In this case, the phenotype curve (in green) does not seem to classify cases and controls in a better way than the permutated ones. Please note that as can be observed in Table 3, the average AUC value of the 80 cycles of phenotypes is 0.535858, while in the case of case of permutated cases and controls it is 0.537543, which means it is 0.31% lower. The column called win subsets indicates the percentage of times when the AUC value of the phenotype is higher than the permutated ones.

Something similar happens in the case of the insulin resistance pathway, where the percentage of the win pathways is 29.75% and the values of the AUC and the average permutated AUC are 0.555483 and 0.556201, respectively (−0.13%). Therefore, in this case, as in the adipocytokine signaling pathway, there does not seem to be any relationship between the two pathways and colorectal cancer. Please also note how in Figure 4 the curve of Phenotype does not seem to be higher than the permutated ones. In addition, for the longevity regulating pathway, whose curve is represented in Figure 5, the situation is similar and no significant influence of those pathways on colorectal cancer can be reported.

Table 3. Pathways under analysis. Total number of single nucleotide polymorphism (SNPs) per pathway (Tot. SNPs), SNPs employed in the 80 iterations by the non-permutated phenotypes (SNPs employed), average area under the receiver operation curve (AUC) obtained in the 80 iterations by the non-permutated phenotypes (AUC), AUC obtained by the permutated phenotypes (AUC perm.), percentage of non-permutated AUC values than are higher than the maximum permutated AUC value (win subsets).

Pathway Name	Tot. SNPs	SNPs Employed	AUC	AUC Perm	Win Subsets
Adipocytokine signalling pathway	752	496	0.535858	0.537543	16.75%
AMPK signaling pathway	1812	462	0.564153	0.551662	89.75%
Apelin signalling pathway	2525	424	0.571761	0.543736	100%
Colorectal cancer pathway	813	423	0.579627	0.565763	100%
Glucagon signalling pathway	1707	487	0.554759	0.552038	82.50%
Huntington's disease	1980	517	0.552436	0.550669	85.00%
Insulin resistance	1574	468	0.555483	0.556201	29.75%
Insulin signalling pathway	1215	451	0.556164	0.552038	96.50%
Longevity regulating pathway	1481	473	0.535285	0.53542	46.75%
Mitochondrial biogenesis	679	438	0.570083	0.552224	100%

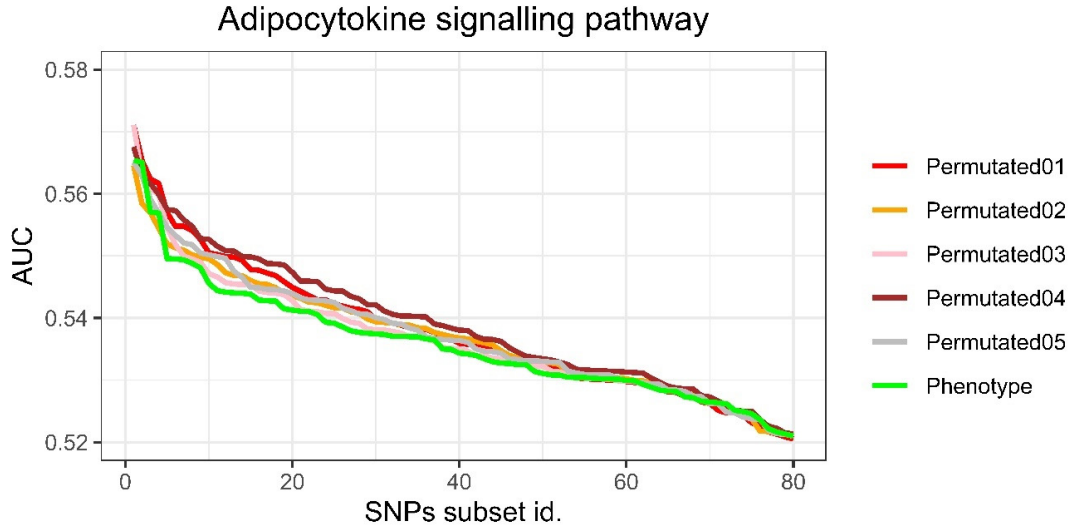

Figure 3. AUC values of the 80 iterations performed for the adipocytokine signaling pathway in the case of cases and controls (phenotype) and five different permutations.

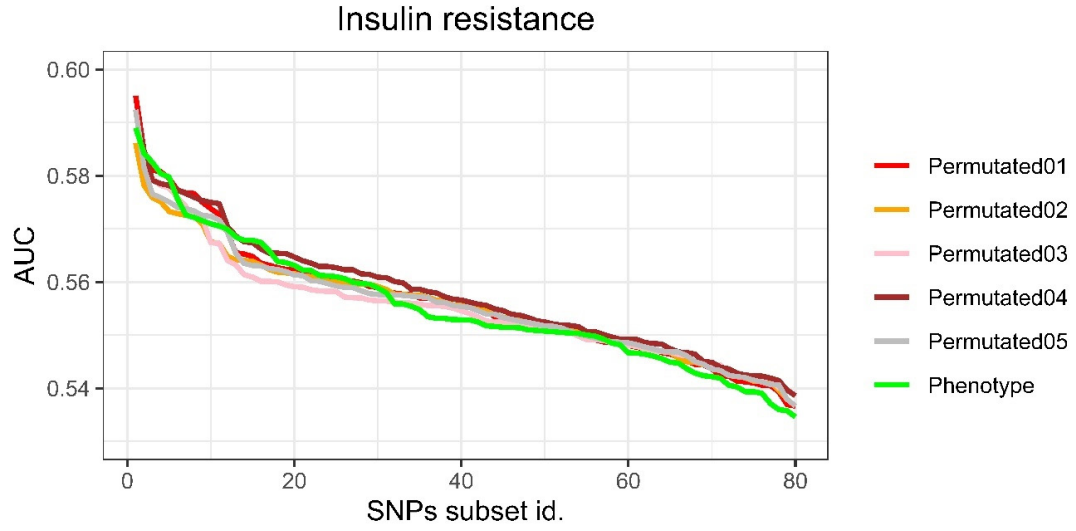

Figure 4. AUC values of the 80 iterations performed for the insulin resistance pathway in the case of cases and controls (phenotype) and five different permutations.

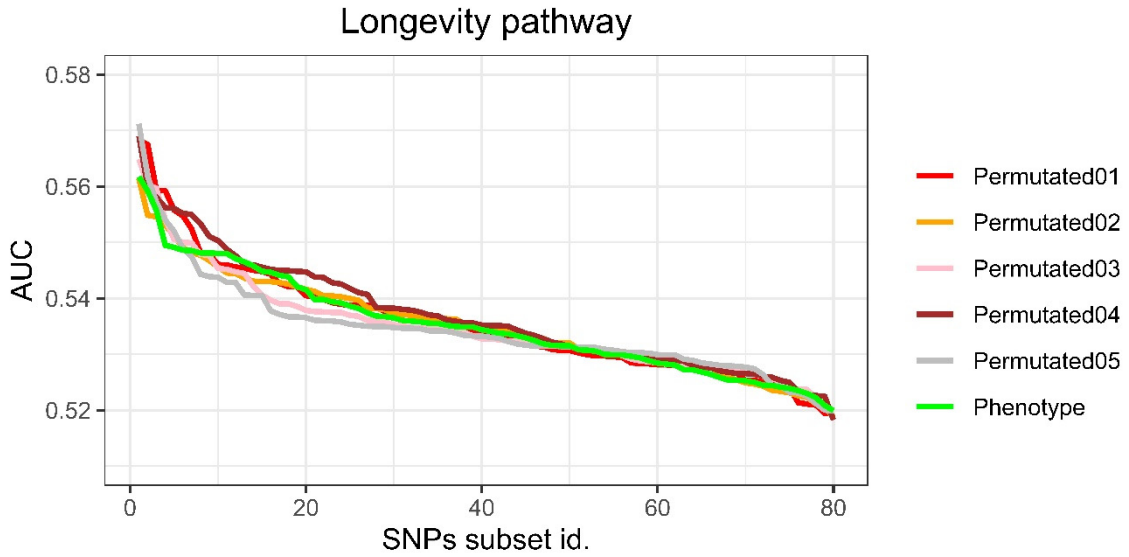

Figure 5. AUC values of the 80 iterations performed for the longevity pathway in the case of cases and controls (phenotype) and five different permutations.

The opposite case applies for apelin signaling, mitochondrial biogenesis and colorectal cancer. In these three cases, the average AUC value obtained for cases and controls are clearly higher than for the permutated phenotypes. In the case of the apelin signaling pathway, it is 5.15% higher in the case of phenotype when compared with the permutated solutions. In the case of mitochondrial biogenesis, it is 3.23% and for the colorectal cancer pathway, the value is 2.45%. In all the cases, the AUC values obtained are higher in the phenotypes than in the permutated cases with 100% of winning cases for apelin signaling pathway, colorectal cancer pathway and mitochondrial biogenesis. Figures 6–8 clearly show how the phenotype curves are higher than the permutated ones. Although in the case of Figure 9, where the AMPK signaling pathway is represented, it does not seem to be as clear as in the three previous cases, the AUC value is 2.26% higher when compared with permutated cases. Please note that in 89.75% of cases the values obtained are higher in the phenotypes than in the permutated cases which, from our point of view, would mean that there is certain influence of this pathway on the trait under analysis.

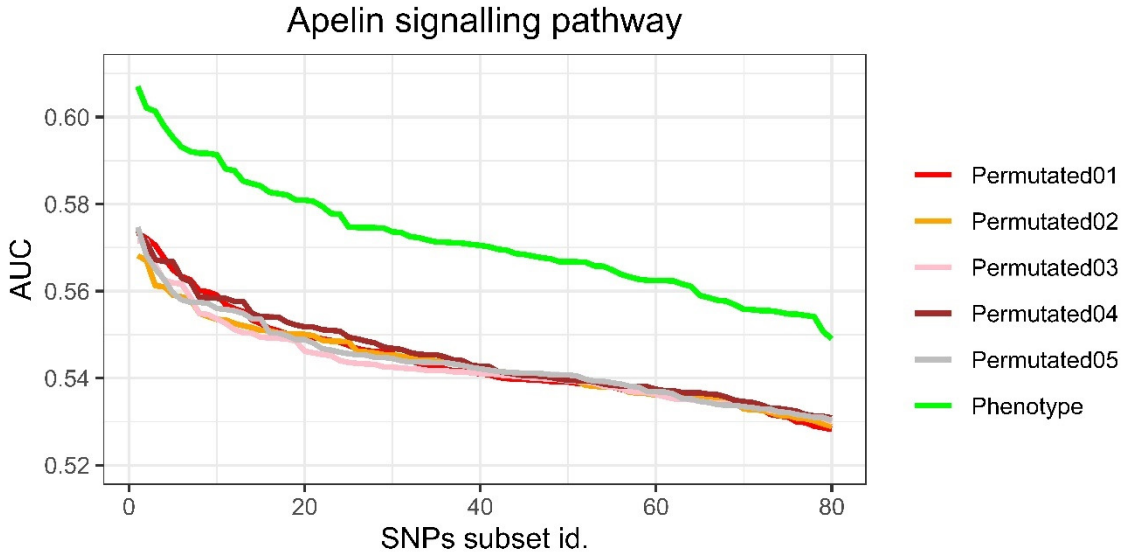

Figure 6. AUC values of the 80 iterations performed for the apelin signaling pathway in the case of cases and controls (phenotype) and five different permutations.

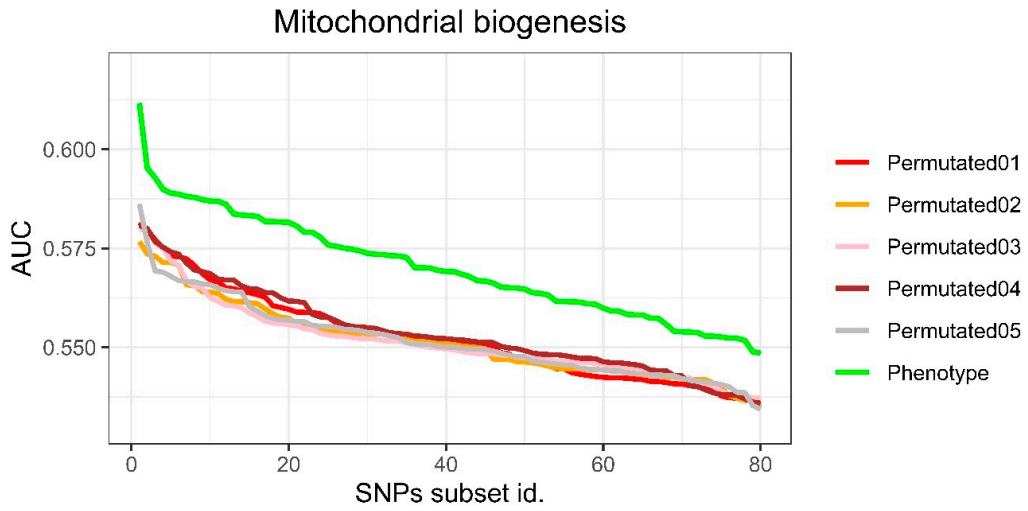

Figure 7. AUC values of the 80 iterations performed for the mitochondrial biogenesis pathway in the case of cases and controls (phenotype) and five different permutations.

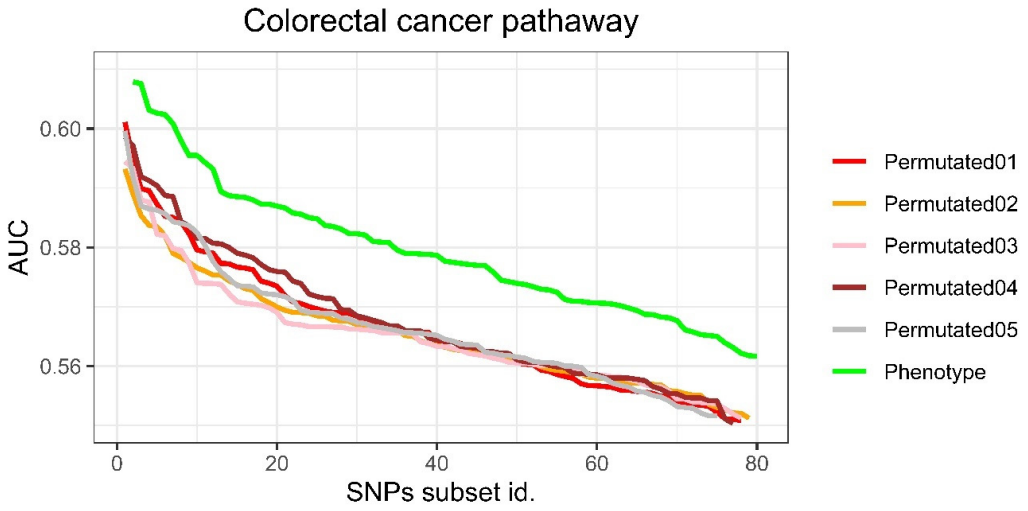

Figure 8. AUC values of the 80 iterations performed for the colorectal cancer pathway in the case of cases and controls (phenotype) and five different permutations.

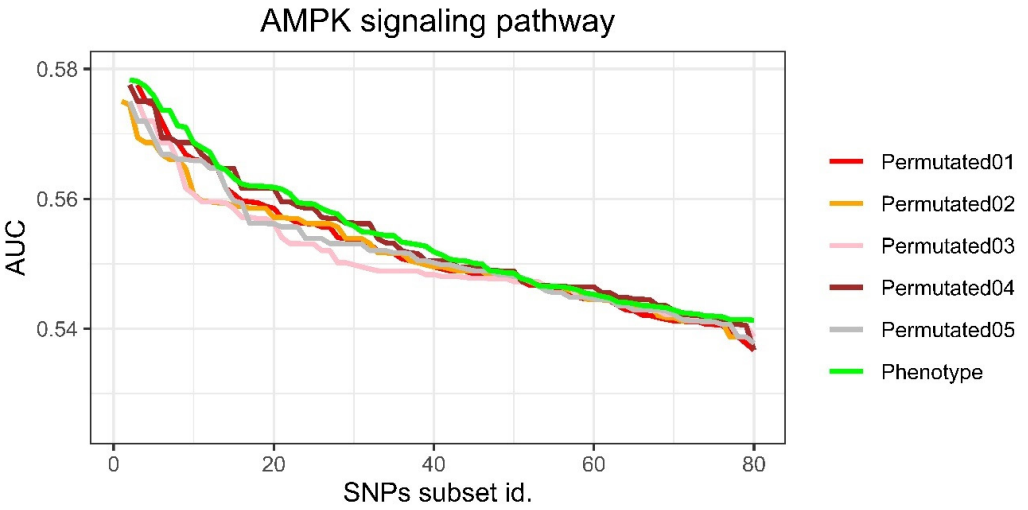

Figure 9. AUC values of the 80 iterations performed for the AMP-activated protein kinase (AMPK) signaling pathway in the case of cases and controls (phenotype) and five different permutations.

The last three pathways under study were glucagon signaling (Figure 10), Huntington's disease (Figure 11) and insulin signaling (Figure 12). In these three pathways, in a similar way to the case of the AMPK signaling pathway, the AUC value of the phenotype is slightly higher than the average value obtained for the permutated ones. Additionally, in most of the cases (from 82.50% to 96.50%), the AUC obtained in the phenotype iteration is higher than in the permutated one.

In summary, taking into account the results obtained with the algorithm proposed in the present research, it can be said that there is a clear relationship linking apelin signaling, colorectal cancer and mitochondrial biogenesis pathways with colorectal cancer. A weak relationship with colorectal cancer was found for AMPK signaling, glucagon signaling, Huntington's disease and insulin signaling pathways. Finally, no relationship with colorectal cancer was found for adipocytokine signaling, insulin resistance or longevity-regulating pathways.

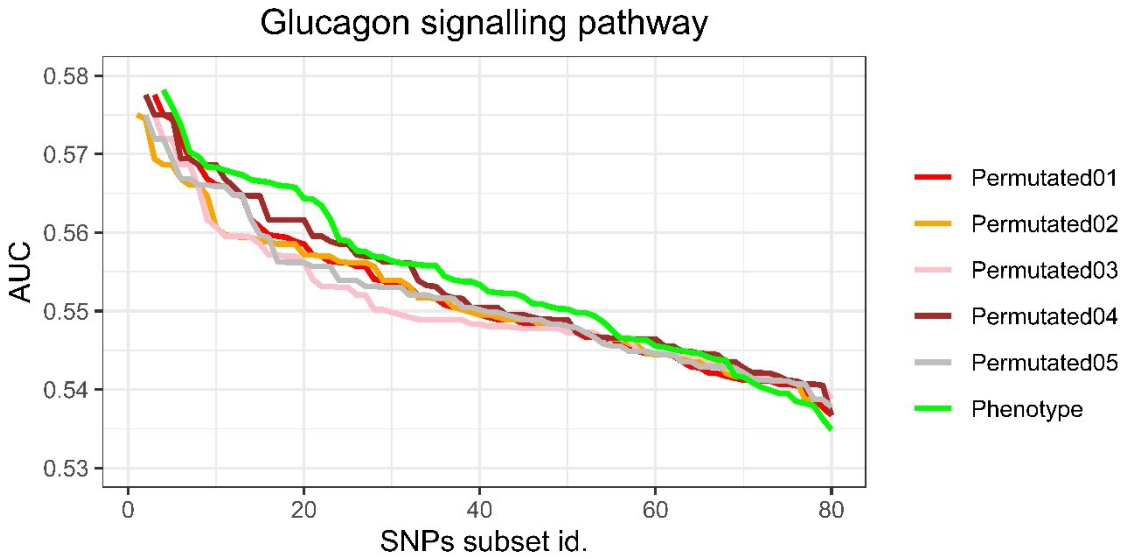

Figure 10. AUC values of the 80 iterations performed for the glucagon signaling pathway in the case of cases and controls (phenotype) and five different permutations.

Figure 11. AUC values of the 80 iterations performed for the Huntington's pathway in the case of cases and controls (phenotype) and five different permutations.

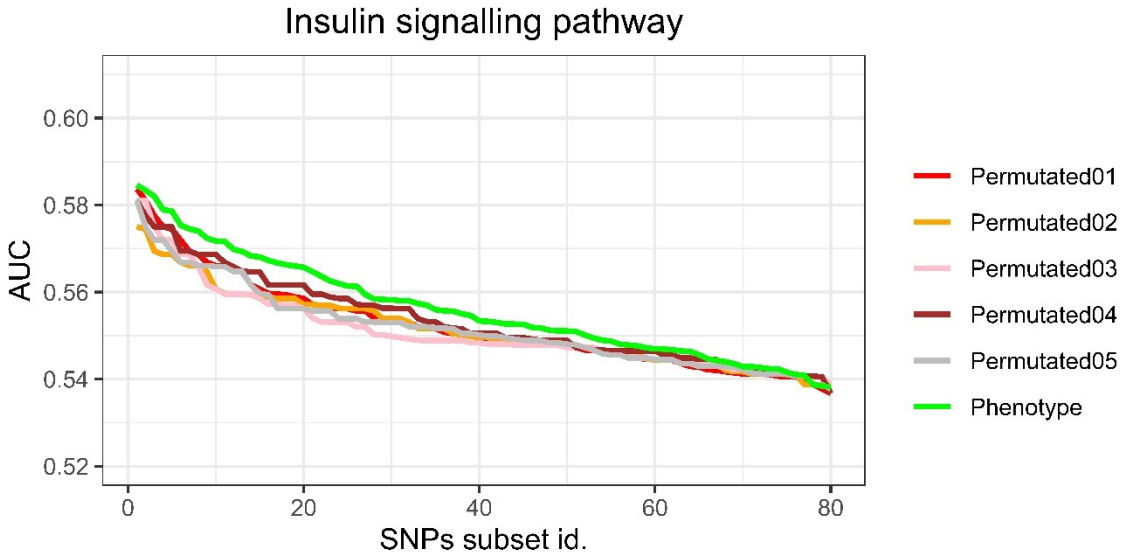

Figure 12. AUC values of the 80 iterations performed for the insulin signaling pathway in the case of cases and controls (phenotype) and five different permutations.

4. Discussion

From the authors' point of view, the results obtained are coherent with previous results to be found in existing literature. Regarding the AUC values, it must be stated that although they would seem to be low, they are in line with other research [43]. In the case of the AMPK signaling pathway, previous studies found that AMPK promotes the survival of colorectal cancer stem cells [44]. This study found that colorectal cancer stem cells show a higher level of antioxidant genes and have a lower level of reactive oxygen species than non-colorectal cancer stem cells. This would be because the colorectal cancer stem cells also possess more mitochondria mass and show higher mitochondrial activity. In the case of this study, higher AMP-activated protein kinase (AMPK) activity was observed in these colorectal cancer stem cells.

Another study that included patients with stage II or III of colorectal cancer, amongst whom the 5-year survival rate was between 50% and 87%, found that the AMPK encoded in gene α1 was overexpressed in patients who suffered from colorectal cancer. For its authors, the AMPK encoded in gene α1 regulate the glutathione reductase (GSR) phosphorylation, possibly through residue Thr507, which enhances its activity. They suggested the suppression of AMPK expression in gene α1 by using nano-sized polymeric vector to induce a favorable therapeutic effect.

In the case of the apelin signaling pathway, which was also found to be relevant to colorectal cancer, there are some studies linking it to colorectal cancer [45]. Apelin is an endogenous ligand of the apelin receptor (APJ), a seven-transmembrane G protein-coupled receptor [45]. It can be found in the brain and also in peripheral organs like the heart, the lungs, blood vessels, and adipose tissue. It is involved in regulating cardiac and vascular function, heart development, and vascular smooth muscle cell proliferation. According to previous research, apelin is not only related to colorectal cancer, but also to others like lung cancer, gastroesophageal, hepatocellular carcinoma, prostate cancer, endometrial cancer, oral squamous cell carcinoma, brain cancer, and tumor neoangiogenesis. This means that Apelin/APJ may be a potential anticancer therapeutic target. A study suggested that the APJ receptor antagonist F13A significantly reduced cellular proliferation [46]. Another study [47] found that Apelin receptor is co-expressed in colorectal cancer cell lines and its activation leads to adenylyl cyclase inhibition and Akt phosphorylation. For the authors of that research, apelin and its receptor might be co-expressed in the tumor compartment where this co-expression would underlie a constitutive activation of apelin signaling and create a functional autocrine loop. It was the first study that

reported that apelin peptide is highly expressed in human colon adenomas and tumors [47]. This co-expression was also observed in several colorectal cancer cell lines. In the LoVo cell line, quantitative real-time polymerase chain reaction (qRT-PCR) experiments and apelin-induced Akt phosphorylation confirmed the concomitant expression of both ligand and receptor. In addition, apelin behaved as an anti-apoptotic peptide, by reversing caspase activation and poly ADP ribose polymerase protein (PARP) degradation induced by theMG132 proteasome inhibitor. Another study [48] that measured apelin, and its receptor mRNA, and protein expression levels in tumor tissue of 56 surgically treated colorectal adenocarcinoma patients and compared them with 27 healthy controls, found that serum levels of apelin and its receptor were increased in colorectal cancer patients in comparison to controls, which leads to the conclusion that apelin could be an important factor in the progression of colorectal carcinoma. The finding of the colorectal cancer pathway as being significant by our algorithm is not surprising, as it can be considered as the pathway of reference.

Mitochondria are semiautonomous organelles that participate in energy metabolism, free radical production, and apoptosis. Apart from the nucleus, the mitochondrion is the only cellular organelle that contains its own genome and genetic machinery [49,50]. Mitochondrial biogenesis is an essential process by which new mitochondria are obtained, and is one which requires coordination between the nuclear and mitochondrial genomes [51]. Mitochondria, as well as most of the processes related to them, are closely linked to the genesis of cancer [52]. For this reason, it is essential to study mitochondrial biogenesis, as well as to find out what happens to these organelles during tumor processes. The progression of CRC in humans is closely linked to mitochondrial alteration, increased production of free mitochondrial oxide radicals, and oxidative stress [53].

An article in a literature review article [54], aimed at evaluating whether increased or decreased peroxisome proliferator-activated receptor gamma coactivator 1-α (PPARGC1A or PGC1α) expression affects the development of colorectal cancer, found that an altered expression of PGC1α modifies colorectal cancer risk and mitochondrial biogenesis is regulated by PGC1α. According to this study, it seems plausible that the proposed algorithm found a relationship between the mitochondrial biogenesis pathway and colorectal cancer.

Glucagon increases the production of glucose by increasing glycogenolysis and gluconeogenesis in the liver, and by reducing glycogenesis and glycolysis. The release of glucagon in response to food consumption depends on the type of meal that has been eaten. If a meal is rich in carbohydrates, blood glucagon levels fall to prevent an undue rise in the level of circulating glucose. Conversely, when a protein-rich meal is eaten, the blood glucagon level rises. Nowadays cancer is known to be one of the major causes of death in diabetic patients, and an association between antidiabetic drugs and the risk of cancer has been reported [55]. Glucagon is nowadays recognized as a pivotal factor implicated in the pathophysiology of diabetes. A recent study has found [56] expression of the glucagon receptor in colon cancer cell lines and in colon cancer tissue obtained from patients. According to this study, glucagon significantly promoted colon cancer cell growth. Molecular assays showed that glucagon acted as an activator of cancer cell growth through deactivation of AMPK and activation of mitogen-activated protein kinase (MAPK). Another study [57] found the relationship between glucagon signaling pathway and endometrial cancer.

A study published in 2002 [58] stated that Huntington's disease provides clues about cancer and that it would be a marker of certain cancers like colorectal cancer. It can be said that, in general, people with Huntington's disease have been observed to have lower rates of cancers [59] and although the relationship of Huntington's disease with prostate cancer has been reported [60], no similar study has been found for colorectal cancer.

The insulin signaling pathway is another of the pathways where a moderate relationship with colorectal cancer was found. It has been reported that the modification in the individual values of plasma insulin levels due to diet may affect the risk of suffering from colorectal cancer [61]. A similar result was found by other researchers in a study performed

with a sample of postmenopausal women [62]. Although there are many studies in this line [63,64] and it is known that genetic variants in metabolic signaling pathways may interact with lifestyle factors such as dietary fatty acids influencing colorectal cancer risk, these interrelated pathways are not fully understood [65].

As mentioned before, the proposed algorithm did not find any relationship linking adypocitokine signaling, insulin resistance and longevity-regulating pathways with colorectal cancer. In the case of the adipocytokine signaling pathway, no relationship has been found in the existing literature with colorectal cancer. However, its relationship with atherosclerosis, diabetes and breast cancer [66] has been reported. In the case of insulin resistance, as far as the authors know, the relationship is not clear enough at this point in time [67]. Finally, in the case of the longevity-regulating pathway, there were no studies found linking it with colorectal cancer.

5. Conclusions

This paper presents a novel method called GASVeM, which is based on two well-known machine learning methodologies—genetic algorithms, and support vector machines. Although the results achieved appear promising, as usually happens in machine learning methodologies applied to GWAS, it is difficult to find a direct biological link between the SNPs involved in the results and the trait under study. In spite of this, through studying existing literature it has been possible for the authors to find previous well-known relationships between the relevant pathways and the trait under analysis.

From the authors' point of view and due to a lack of a machine learning gold standard for GWAS analysis, the present method could be of interest for future GWAS. In this direction, based on the results obtained, it would be interesting, on the one hand, to work on studying the classification capacity when information from several pathways is combined and, on the other hand, to try to replicate the results obtained in other databases and in the analysis of other pathologies, to validate the usefulness of the method under different conditions of use.

We also agree with those authors that consider that we are in the infancy of the use of machine learning in GWAS [68] as we are still quite a long way from achieving gold standard methods producing consistently validated biological insights. In addition, we would like to highlight the characteristics of this kind of database, whereby a high number of SNPs (columns) when compared with the number of cases (rows) cause a kind of problem with GWAS that is difficult to deal with from a machine learning point of view and that, in the case of the present research, is present in the need for an a priori SNPs selection based on pathways. It is the authors' opinion that this problem has a great impact on the reproducibility of results when the same algorithm is applied to a different database.

Finally, it must be said that we are aware that the translation of the results obtained with this method to a population-based clinical practice to carry out a personalization of interventions based on genomic data still requires further steps to be taken before the selection capacity can be refined. However, we consider that the method has demonstrated, as was our objective, a good ability to discriminate which pathways are associated with the event and which are not, through the choice of a limited set of SNPs. We consider the ability to classify individuals to be a second step to be taken in the development of the model, through lines of study such as the inclusion of several pathways in the model.

Author Contributions: Conceptualization, F.D.D., V.M., V.M.S. and F.S.L.; data curation, F.M.-N. and A.J.M.d.l.T.; formal analysis, V.M.; methodology, F.S.L. and F.D.D.; project administration, V.M.S.; resources, F.M.-N. and A.J.M.d.l.T.; software, F.D.D. and F.S.L.; validation, F.D.D.; visualization, F.S.L.; writing—original draft, F.D.D. and F.S.L.; writing—review and editing, F.D.D., F.S.L., V.M., F.M.-N., A.J.M.d.l.T. and V.M.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by Agency for Management of University and Research Grants (AGAUR) of the Catalan Government, grant number 2017SGR723, Instituto de Salud Carlos III, co-funded by FEDER funds –a way to build Europe– grants and Spanish Association Against Cancer (AECC) Scientific Foundation grant GCTRA18022MORE.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Institutional Review Board of Instituto Municipal de Asistencia Sanitaria de Barcelona (Spain) with protocol code 2008/3123/I on date 3rd of September 2008 and also approved the 29th of May 2009 by the Ethical Committe of Leon Hospital (Spain) without any protocol number.

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

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy.

Acknowledgments: The authors would like to thank Anthony Ashworth for his revision of the English grammar and spelling in the manuscript. We thank CERCA Programme, Generalitat de Catalunya for institutional support.

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

References

1. Venter, J.C. The sequence of the human genome. *Science* **2001**, *291*, 1304–1351. [CrossRef]
2. Gibbs, R.A.; Belmont, J.W.; Hardenbol, P.; Willis, T.D.; Yu, F.L.; Yang, H.M.; Ch'ang, L.Y.; Huang, W.; Liu, B.; Shen, Y.; et al. The International HapMap Project. *Nature* **2003**, *426*, 789–796.
3. Slatkin, M. Linkage disequilibrium—Understanding the evolutionary past and mapping the medical future. *Nat. Rev. Genet.* **2008**, *9*, 477–485. [CrossRef]
4. Appasani, K. *Genome-Wide Association Studies*; Cambridge University Press: Cambridge, UK, 2015.
5. Bergen, S.E.; Petryshen, T.L. Genome-wide association studies of schizophrenia: Does bigger lead to better results? *Curr. Opin. Psychiatry* **2012**, *25*, 76–82. [CrossRef] [PubMed]
6. Frazer, K.A.; Ballinger, D.G.; Cox, D.R.; Hinds, D.A.; Stuve, L.L.; Gibbs, R.A.; Belmont, J.W.; Boudreau, A.; Hardenbol, P.; Leal, S.M. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **2007**, *449*, 851–861. [PubMed]
7. Klein, R.J.; Zeiss, C.; Chew, E.Y.; Tsai, J.Y.; Sackler, R.S.; Haynes, C.; Henning, A.K.; SanGiovanni, J.P.; Mane, S.M.; Mayne, S.T.; et al. Complement factor H polymorphism in age-related macular degeneration. *Science* **2005**, *308*, 385–389. [CrossRef] [PubMed]
8. DeWan, A.; Liu, M.; Hartman, S.; Zhang, S.S.; Liu, D.T.; Zhao, C.; Tam, P.O.; Chan, W.M.; Lam, D.S.; Snyder, M.; et al. HTRA1 promoter polymorphism in wet age-related macular degeneration. *Science* **2006**, *314*, 989–992. [CrossRef] [PubMed]
9. Ziegler, A.; Ghosh, S.; Dyer, T.D.; Blangero, J.; Maccluer, J.; Almasy, L. Introduction to genetic analysis workshop 17 summaries. *Gen. Epidemiol.* **2011**, *35*, S1–S4. [CrossRef] [PubMed]
10. Tabor, H.K.; Risch, N.J.; Myers, R.M. Candidate-gene approaches for studying complex genetic traits: Practical considerations. *Nat. Rev. Genet.* **2002**, *3*, 391–396. [CrossRef]
11. Lippert, C.; Listgarten, J.; Davidson, R.I.; Baxter, S.; Poon, H.; Cadie, C.M.; Heckerman, D. An exhaustive epistatic SNP association analysis on expanded Wellcome Trust data. *Sci. Rep.* **2013**, *3*, 1099. [CrossRef]
12. Ning, C.; Wang, D.; Zhou, L.; Wei, J.; Liu, Y.; Kang, H.; Zhang, S.; Zhou, X.; Xu, S.; Liu, J.F. Efficient multivariate analysis algorithms for longitudinal genome-wide association studies. *Bioinformatics* **2019**, *35*, 4879–4885. [CrossRef] [PubMed]
13. Romagnoni, A.; Jégou, S.; Van Steen, K.; Wainrib, G.; Hugot, J.P. Comparative performances of machine learning methods for classifying Crohn Disease patients using genome-wide genotyping data. *Sci. Rep.* **2019**, *9*, 10351. [CrossRef] [PubMed]
14. Lin, H.; Hargreaves, K.A.; Li, R.; Reiter, J.L.; Wang, Y.; Mort, M.; Cooper, D.N.; Zhou, Y.; Eadon, M.T.; Dolan, M.E.; et al. RegSNPs-intron: A computational framework for predicting pathogenic impact of intronic single nucleotide variants. *Genome Biol.* **2019**, *20*, 254. [CrossRef] [PubMed]
15. Mackay, T.F. Epistasis and quantitative traits: Using model organisms to study gene-gene interactions. *Nat. Rev. Genet.* **2014**, *15*, 22–33. [CrossRef]
16. Artime Ríos, E.; Suárez Sánchez, A.; Sánchez Lasheras, F.; Seguí Crespo, M.M. Genetic algorithm based on support vector machines for computer vision syndrome classification in health personnel. *Neural Comput. Appl.* **2020**, *32*, 1239–1248. [CrossRef]
17. Vilán Vilán, J.A.; Alonso Fernández, J.R.; García Nieto, P.J.; Sánchez Lasheras, F.; de Cos Juez, F.J.; Díaz Muñiz, C. Support Vector Machines and Multilayer Perceptron Networks Used to Evaluate the Cyanotoxins Presence from Experimental Cyanobacteria Concentrations in the Trasona Reservoir (Northern Spain). *Water Resour. Manag.* **2013**, *27*, 3457–3476. [CrossRef]
18. Casteleiro-Roca, J.L.; Jove, E.; Sánchez-Lasheras, F.; Méndez-Pérez, J.A.; Calvo-Rolle, J.L.; de Cos Juez, F.J. Power Cell SOC Modelling for Intelligent Virtual Sensor Implementation. *J. Sens.* **2017**, *2017*, 9640546. [CrossRef]
19. Deisenroth, M.P.; Faisal, A.A.; Cheng, S.O. *Mathematics for Machine Learning*; Cambridge University Press: Cambridge, UK, 2020.
20. Kramer, O. *Genetic Algorithm Essentials*; Springer International Publishing: New York, NY, USA, 2017.
21. Holland, J.H. *Adaptation in Natural and Artificial Systems*; MIT Press: London, UK, 1992.
22. Rechenberg, I. *Evolutionsstrategie*; Holzmann-Froboog: Stuttgart, Germany, 1973.
23. Schwefel, H.P. *Numerical Optimization of Computer Models*; Wiley: Chichester, NY, USA, 1981.
24. Vose, M.D. *The Simple Genetic Algorithm. Foundations and Theory*; The MIT Press: Cambridge, MA, USA, 1999.

25. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer Series in Statistics: New York, NY, USA, 2009.
26. Gondro, C.; van der Werf, J.; Hayes, B. (Eds.) *Genome-Wide Association Studies and Genomic Prediction*; Methods in Molecular Biology; Humana Press: New York, NY, USA, 2013.
27. Marozzi, M. A bi-aspect nonparametric test for the two-sample location problem. *Comput. Stat. Data Anal.* **2002**, *64*, 639–648. [CrossRef]
28. Anderson, M.J.; Legendre, P. An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. *J. Stat. Comput. Sim.* **1999**, *62*, 271–303. [CrossRef]
29. Shipey, B. A permutation procedure for testing the equality of pattern hypotheses across groups involving correlation or covariance matrix. *Stat. Comput.* **2000**, *10*, 253–257. [CrossRef]
30. Ernst, M.D.; Schucany, W.R. A Class of Permutation Tests of Bivariate Interchangeability. *J. Am. Stat. Assoc.* **1999**, *94*, 273–284. [CrossRef]
31. Pesarin, F. Goodness of fit for ordered discrete distributions by resampling techniques. *Metron* **1994**, *52*, 57–71.
32. Marozzi, M. Some remarks about the number of permutations one should consider to perform a permutation test. *Statistica* **2004**, *64*, 193–201.
33. Edgington, E.S. *Randomization Tests*, 3rd ed.; Dekker: New York, NY, USA, 1995.
34. Browning, B.L. PRESTO: Rapid calculation of order statistic distributions and multiple-testing adjusted P-values via permutation for one and two-stage genetic association studies. *BMC Bioinform.* **2008**, *9*, 309. [CrossRef] [PubMed]
35. Wolpert, D.H.; Macready, W.G. No free lunch theorems for optimization. *IEEE Trans. Evol. Comput.* **1997**, *1*, 67–82. [CrossRef]
36. De Landgraaf, W.A.; Eiben, A.E.; Nannen, V. Parameter calibration using meta-algorithms. In Proceedings of the 2007 IEEE Congress on Evolutionary Computation, Singapore, 25–28 September 2007.
37. Bäck, T.; Schütz, M. Intelligent mutation rate control in canonical genetic algorithms. In *Foundation of Intelligent Systems, Proceedings of the 9th International Symposium, ISMIS '96, Zakopane, Poland, 9–13 June 1996*; Springer: Berlin/Heidelberg, Germany, 1996; pp. 158–167.
38. Deng, S.; Perez-Cardona, J.; Huang, A.; Yih, Y.; Thompson, V.S.; Reed, D.W.; Jin, H.; Sutherland, J.W. Applying design of experiments to evaluate economic feasibility of rare-earth element recovery. *Procedia CIRP* **2020**, *90*, 165–170. [CrossRef]
39. Wang, C.N.; Dang, T.T.; Nguyen, N.A.T. A Computational Model for Determining Levels of Factors in Inventory Management Using Response Surface Methodology. *Mathematics* **2020**, *8*, 1210. [CrossRef]
40. Kanehisa, M.; Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **2000**, *28*, 27–30. [CrossRef]
41. Kanehisa, M. Toward understanding the origin and evolution of cellular organisms. *Protein Sci.* **2019**, *28*, 1947–1951. [CrossRef]
42. Kanehisa, M.; Furumichi, M.; Sato, Y.; Ishiguro-Watanabe, M.; Tanabe, M. KEGG: Integrating viruses and cellular organisms. *Nucleic Acids Res.* **2021**, *49*, D545–D551. [CrossRef]
43. Thomas, M.; Sakoda, L.C.; Hoffmeister, M.; Rosenthal, E.A.; Lee, J.K.; van Duijnhoven, F.J.B.; Platz, E.A.; Wu, A.H.; Dampier, C.H.; de la Chapelle, A.; et al. Genome-Wide Modeling of Polygenic Risk Score in Colorectal Cancer Risk. *Am. J. Hum. Genet.* **2020**, *107*, 432–444. [CrossRef]
44. Guo, B.; Han, X.; Tkach, D.; Huang, S.G.; Zhang, D. AMPK promotes the survival of colorectal cancer stem cells. *Anim. Models Exp. Med.* **2018**, *1*, 134–142. [CrossRef] [PubMed]
45. Yang, Y.; Lv, S.Y.; Ye, W.; Zhang, L. Apelin/APJ system and cancer. *Clin. Chim. Acta* **2016**, *457*, 112–116. [CrossRef] [PubMed]
46. Mughal, A.; O'Rourke, S.T. Vascular effects of apelin: Mechanisms and therapeutic potential. *Pharmacol. Ther.* **2018**, *190*, 139–147. [CrossRef] [PubMed]
47. Picault, F.X.; Chaves-Almagro, C.; Projetti, F. Tumour co-expression of apelin and its receptor is the basis of an autocrine loop involved in the growth of colon adenocarcinomas. *Eur. J. Cancer* **2014**, *50*, 663–674. [CrossRef] [PubMed]
48. Podgórska, M.; Diakowska, D.; Pietraszek-Gremplewicz, K.; Nienartowicz, M.; Nowak, D. Evaluation of Apelin and Apelin Receptor Level in the Primary Tumor and Serum of Colorectal Cancer Patients. *J. Clin. Med.* **2019**, *8*, 1513. [CrossRef]
49. Permuth-Wey, J.; Chen, Y.A.; Tsai, Y.Y.; Chen, Z. Inherited Variants in Mitochondrial Biogenesis Genes May Influence Epithelial Ovarian Cancer Risk. *Cancer Epidemiol. Prev. Biomark.* **2011**, *20*, 1131–1145. [CrossRef]
50. Baar, K.; Song, Z.; Semenkovich, F.C.; Jones, T.E.; Han, D.H.; Nolte, L.A.; Ojuca, E.O. Skeletal muscle overexpression of nuclear respiratory factor 1 increases glucose transport capacity. *FASEB J.* **2003**, *17*, 1666–1673. [CrossRef]
51. Blesa, J.R.; Prieto Ruiz, J.A.; Abraham, B.A.; Harrison, B.L.; Hedge, A.A.; Hernández Yago, J. NRF-1 is the major transcription factor regulating the expression of the human TOMM34 gene. *Biochem. Cell Biol.* **2008**, *86*, 46–56. [CrossRef]
52. Skonieczna, K.; Malyarchuk, B.A.; Grzybowski, T. The landscape of mitochondrial DNA variation in human colorectal cancer on the background of phylogenetic knowledge. *Biochim. Biophys. Acta* **2012**, *1825*, 153–1599. [CrossRef]
53. Sanchez Pino, M.J.; Moreno, P.; Navarro, A. Mitochondrial dysfunction in human colorectal cancer progression. *Front. Biosci.* **2007**, *12*, 1190–1199. [CrossRef] [PubMed]
54. Alonso Molero, J.; González Donquiles, C.; Fernández Villa, T.; de Souza Teixeira, F.; Vilorio Marqués, L.; Molina, A.J.; Martín, V. Alterations in PGC1α expression levels are involved in colorectal cancer risk: A qualitative systematic review. *BMC Cancer* **2017**, *17*, 731. [CrossRef] [PubMed]
55. Yagi, T.; Kubota, E.; Koyama, H.; Tanaka, T.; Kataoka, H.; Imaeda, K.; Joh, T. Glucagon promotes colon cancer cell growth via regulating AMPK and MAPK pathways. *Oncotarget* **2018**, *9*, 10650–10664. [CrossRef] [PubMed]

56. Wu, Z.; Liu, Z.; Ge, W.; Shou, J.; You, L.; Pan, H.; Han, W. Analysis of potential genes and pathways associated with the colorectal normal mucosa-adenoma-carcinoma sequence. *Cancer Med.* **2018**, *7*, 2555–2566. [CrossRef]
57. Kanda, R.; Hiraike, H.; Wada-Hiraike, O.; Ichinose, T.; Nagasaka, K.; Sasajima, Y.; Ryo, E.; Fujii, T.; Osuga, Y.; Ayabe, T. Expression of the glucagon-like peptide-1 receptor and its role in regulating autophagy in endometrial cancer. *BMC Cancer* **2018**, *18*, 657. [CrossRef] [PubMed]
58. Kerr, C. Huntington's disease provides cancer clues. *Lancet Oncol.* **2002**, *3*, 518. [CrossRef]
59. McNulty, P.; Pilcher, R.; Ramesh, R.; Necuiniate, R.; Hughes, A.; Farewell, D.; Holmans, P.; Jones, L.; REGISTRY Investigators of the European Huntington's Disease Network. Reduced Cancer Incidence in Huntington's Disease: Analysis in the Registry Study. *J. Huntingt. Dis.* **2018**, *7*, 209–222.
60. Huang, Y.F.; Yeh, H.Y.; Soo, V.W. Inferring drug-disease associations from integration of chemical, genomic and phenotype data using network propagation. *BMC Med. Genom.* **2013**, *6*, S4. [CrossRef]
61. Pechlivanis, S.; Pardini, B.; Bermejo, J.L.; Wagner, K.; Naccarati, A.; Vodickova, L.; Novotny, J.; Hemminki, K.; Vodicka, P.; Försti, A. Insulin pathway related genes and risk of colorectal cancer: INSR promoter polymorphism shows a protective effect. *Endocr. Relat. Cancer* **2007**, *14*, 733–740. [CrossRef]
62. Jung, S.Y.; Rohan, T.; Strickler, H.; Bea, J.; Zhang, Z.F.; Ho, G.; Crandall, C. Genetic variants and traits related to insulin-like growth factor-I and insulin resistance and their interaction with lifestyles on postmenopausal colorectal cancer risk. *PLoS ONE* **2017**, *12*, e0186296. [CrossRef]
63. Poloz, Y.; Stambolic, V. Obesity and cancer, a case for insulin signaling. *Cell Death Dis.* **2015**, *6*, e2037. [CrossRef] [PubMed]
64. Lohmann, A.E.; Goodwin, P.J.; Chlebowski, R.T.; Pan, K.; Stambolic, V.; Dowling, R.J.O. Association of Obesity-Related Metabolic Disruptions with Cancer Risk and Outcome. *J. Clin. Oncol.* **2016**, *34*, 4249–4255. [CrossRef]
65. Jung, S.Y.; Zhang, Z.F. The effects of genetic variants related to insulin metabolism pathways and the interactions with lifestyles on colorectal cancer risk. *Menopause* **2019**, *26*, 771–780. [CrossRef]
66. Li, J.; Han, X. Adipocytokines and breast cancer. *Curr. Probl. Cancer* **2018**, *42*, 208–214. [CrossRef] [PubMed]
67. Tabung, F.K.; Wang, W.; Fung, T.T.; Smith-Warner, S.A.; Keum, N.; Wu, K.; Fuchs, C.S.; Hu, F.B.; Giovannucci, E.L. Association of dietary insulinemic potential and colorectal cancer risk in men and women. *Am. J. Clin. Nutr.* **2018**, *108*, 363–370. [CrossRef] [PubMed]
68. Nicholls, H.L.; John, C.R.; Watson, D.S.; Munroe, P.B.; Barnes, M.R.; Cabrera, C.P. Reaching the End-Game for GWAS: Machine Learning Approaches for the Prioritization of Complex Disease Loci. *Front. Genet.* **2020**, *11*, 350. [CrossRef] [PubMed]