Computers in Biology and Medicine 179 (2024) 108924

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

DiGAS: Differential gene allele spectrum as a descriptor in genetic studies

Antonino Aparo [a,d], Vincenzo Bonnici [b], Simone Avesani [a], Luciano Cascione [c], Rosalba Giugno [a,*], for the Alzheimer's Disease Neuroimaging Initiative [1]

[a] University of Verona, Strada le Grazie, 15, Verona, 37134, Italy
[b] University of Parma, Parco Area delle Scienze, 53/A, Parma, 43124, Italy
[c] Institute of Oncology Research (IOR), Via Francesco Chiesa 5, Bellinzona, 6500, Switzerland
[d] Research Center LURM (Interdepartmental Laboratory of Medical Research), University of Verona, Ple. L.A. Scuro 10, Verona, 37134, Italy

ARTICLE INFO

MSC:
00-01
99-00

Keywords:
Genomic variations
Alzheimer's disease
Classification
Gene allele

ABSTRACT

Diagnosing individuals with complex genetic diseases is a challenging task. Computational methodologies exploit information at the genotype level by taking into account single nucleotide polymorphisms (SNPs) leveraging the results of genome-wide association studies analysis to assign a statistical significance to each SNP. Recent methodologies extend such an approach by aggregating SNP significance at the genetic level to identify genes that are related to the condition under study. However, such methodologies still suffer from the initial SNP analysis limitations. Here, we present DiGAS, a tool for diagnosing genetic conditions by computing significance, by means of SNP information, directly at the complex level of genetic regions. Such an approach is based on a generalized notion of allele spectrum, which evaluates the complete genetic alterations of the SNP set belonging to a genetic region at the population level. The statistical significance of a region is then evaluated through a differential allele spectrum analysis between the conditions of individuals belonging to the population. Tests, performed on well-established datasets regarding Alzheimer's disease, show that DiGAS outperforms the state of the art in distinguishing between sick and healthy subjects.

1. Introduction

All human beings are on average 99.9% identical in their genetic makeup. However, differences in the remaining 0.1% can significantly affect human health. In human diseases, these variations can manifest as variants involving single nucleotide changes called single nucleotide polymorphisms (SNPs) and along this text also as single variants, or non-SNP genetic variants such as insertions, deletions or larger genomic rearrangements. Among these, SNPs represent the most abundant genetic variation in the human genome, occurring once every 300 base pairs throughout the genome [1]. The primary focus on SNPs in genetic analysis is justified by their abundance, wide genomic coverage, heritability, functional impact, relevance in population studies, and clinical applications. SNPs, characterized by single nucleotide substitution, follow Mendelian inheritance patterns and contribute to the heritability of diseases and traits [2]. The analysis of one or multiple SNPs within a gene or in intergenic non-coding regions allows researchers to identify the underlying mechanisms of diseases, gaining insights into the assessment of disease risk, and develop targeted therapies for more personalized approaches [3,4]. For instance, a specific single nucleotide polymorphism in the APOE gene has been shown to contribute to the development of Alzheimer's disease [5,6]. Likewise a number of SNPs within several immune response genes were found to influence the individual susceptibility to autoimmune infectious diseases [3] and, additionally, the identification of rare genomic variations has enabled the development of targeted therapies for cystic fibrosis [4].

Genome-Wide Association Study (GWAS) is a well-established methodology for identifying single variants associated with disease risk [7,8]. GWAS allows testing of hundreds of thousands of SNPs across entire genomes to find those that are statistically associated with a specific disease outcome. Although the analysis of individual SNPs has proven useful in the identification of different disease-susceptibility variants, GWAS testing of millions of variants is however constrained by multiple hypothesis testing [9], which increases chance of yielding false-positive results, inhibiting the validity of the analysis results.

* Corresponding author.
E-mail address: rosalba.giugno@example.com (R. Giugno).
URL: https://infomics.github.io/InfOmics/ (R. Giugno).

[1] Data used in preparation of this article were obtained from the Alzheimer2019s Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

https://doi.org/10.1016/j.compbiomed.2024.108924
Received 1 February 2024; Received in revised form 16 July 2024; Accepted 17 July 2024
Available online 26 July 2024

Indeed, the capability of GWAS to detect SNPs with small effects that are really linked to the outcome, also called causal SNPs, is limited and individual SNPs genotyped on GWAS platforms commonly exhibit only modest effects at phenotypic level. This could be the case when there are typed SNPs that are in linkage disequilibrium (LD) with the causal SNP, meaning SNPs belonging to different but non-randomly associated loci. Therefore, when GWAS analysis is performed, the SNPs in LD with the causal SNP show only moderate effects because serves as an imperfect surrogate for the causal SNP. Given the high LD rate of many typed SNPs with the causal SNP, it could be beneficial to consider in the analysis the simultaneous effect of multiple SNPs in order to capture the effective cause of a phenotypic condition, thus increasing the significance of the results [10].

Highly effective methods for predicting the impact of SNPs are also those based on deep learning models such as DeepSEA [11] and DanQ [12]. These models leverage convolutional and recurrent neural networks to predict the functional effects of non-coding single variants directly from DNA sequences. DeepSEA uses DNA sequences as input and integrates additional features derived from ChIP-seq data, instead DanQ extends this approach by combining convolutional layers with bi-directional long short-term memory (LSTM) units to capture dependencies in the data. Although these models demonstrate high performance in terms of predictive power of specific single variants, they require large amounts of labeled training data and significant computational resources. Moreover, single variants analysis only accounts for the marginal effects of each variant, without considering the epistatic interactions that predispose to disease with larger effects [13,14].

Analyzing sets of SNPs instead of single variants across the entire genome may provide more robust and biologically meaningful results. This approach leverages information from multiple SNPs grouped according to the biological context of their respective regions. The power and relevance of their analyses can be enhanced by focusing on candidate genomic regions, such as promoter regions or tissue-specific genes, or prioritizing candidate genes known to play significant roles in specific pathways Moreover, this approach also allows to reduce the number of possible tests, improve the statistical power, and identify novel loci without increasing sample sizes or collecting new data. Thus, the chance to achieve significant results increases when biological evidence and statistical significance are combined.

In this perspective, in 2011, Wu et al. proposed SKAT [15], a test to evaluate SNP sets applying a logistic kernel-machine regression framework to measure the combined effect of independent SNPs. SNPs are grouped according to genomic features such as genes or haplotype blocks considering SNP-sets as potential regulatory regions, reducing the number of multiple comparisons. The goal of SNP set analysis is to test the global null hypothesis of whether any of the SNPs are related to the outcome while adjusting for the additional covariates. Besides SKAT, other SNP sets tests have become increasingly important in analyzing the association problem at the gene level through the computation of gene-level p-values or gene scores.

In minSNP [16–18], the SNP with the smallest p-value is used as a representative of the entire gene.

Alternatively, in permSNP [19–22] an empirical p-value for a SNP set is determined by recomputing the p-values of individual SNPs using a permuted dataset. The SNP set's p-value is then calculated as the number of times where the average p-value of the observed SNPs is lower than the p-values obtained from the permuted data.

A similar empirical p-value is also computed in VEGAS [23] where a multivariate normal distribution is used to correct for uneven LD distribution between SNPs. Alternatively, Pegasus [24] employs a chi-squared distribution to capture LD between SNPs at gene level. However, all these methods inherit from GWAS the issue of assigning significance to each SNP in a SNP analysis before grouping SNPs into sets. Table 1 summarizes the main characteristics of these approaches along with their limitations.

In this context, we introduce DiGAS, a tool implementing an innovative computational approach for the identification of genomic elements, ranging from individual exons to entire genomic regions, likely associated with a given phenotypic condition, such as diseases, treating them as potential causal factors. DiGAS introduces a novel genomic information descriptor named the "generalized allele spectrum". This descriptor is built upon the allele frequency spectrum, which captures allele frequencies within a defined group of loci, specifically SNPs. The allele spectrum combines the frequency of single alleles into an unique vector of allele frequencies.

In contrast to the allele spectrum, the novel descriptor takes into account the complete set of SNPs of a region at once, allowing to compute frequency at the genomic region level rather than at the SNP level. We define the Differential Generalized Allele Spectrum as the set of significant differences in the frequency allele spectra between two sample conditions. The proposed methodology (i) recognizes genetic regions critical for a given phenotype, and (ii) builds a set of features for supervised classification purposes.

DiGAS was tested on a case study of Alzheimer's disease(AD), a perfect scenario where, considering the collective influence of multiple SNPs, we can gain a better understanding of the genetic architecture underlying the disease and potentially identify more comprehensive sets of genetic markers associated with its progression [25,26]. AD is a progressive neurodegenerative disease that induces a slow and inevitable degeneration of brain functions. Up to date, AD has no cure, and it represents a challenge at the forefront of biomedical research [27]. Genetic factors play a significant role in the development and progression of AD, with variations in specific genes increasing the risk of developing the condition. In studying Alzheimer's disease, it has been observed that a particular SNP may be present and associated with the condition in one affected individual but not in another affected one meaning that the presence or absence of a single specific SNP is not sufficient to determine the disease status or predict its occurrence. Instead, AD is influenced by the combined effect of multiple SNPs that may vary between individuals. Each individual may have a unique combination of genetic variations, including different SNPs, that contribute to their susceptibility or resilience to the disease [28,29].

We collected genetic data of AD patients from the Alzheimer's Disease Neuroimaging Initiative (ADNI) [30] and used DiGAS to identify key sets of SNPs that differentiate between healthy and AD individuals. DiGAS was then compared with SKAT, which served as benchmark method in the context of multiple SNPs analyses. Results show that DiGAS outperforms SKAT in the identification of predictive SNP sets for the classification task.

To further evaluate the effectiveness of DiGAS on different phenotypic conditions, we tested it on openly available data collected from the Parkinson's Progression Markers Initiative (PPMI) [31], which was obtained upon request. The results confirmed that DiGAS exhibit superior performance compared to SKAT in the identification of predictive SNP sets, reducing also significantly both the computational time and memory usage.

DiGAS is implemented in Python and the open-source software is available for both Windows and Unix systems at the following GitHub repository: https://github.com/InfOmics/DiGAS.

In what follows we present a detailed description of the DiGAS methodology and the results obtained by testing the tool on the AD case study. Section 2.1 introduces the main methodological aspects of the proposed approach. Section 2.2 describes the datasets used for the evaluation of the proposed model. Finally, the results in the form of a supervised classification problem, are reported in Section 3.

2. Material and methods

In this section, we present the proposed methodology, DiGAS, along with details about the data used for testing and the validation approach.

Table 1
A summary of the most commonly used SNP sets methods and limitations.

Methods	Description	Limitations
minSNP [16–18]	Computes a gene score based on the smallest SNP *p*-value observed within the gene in a GWAS.	Biases may occur as longer genes tend to have lower gene scores.
permSNP [19–22]	Involves permuting case-control labels in genotype data,recalculating SNP p-values, and computing empirical gene p-values using the observed and permuted data.	Computationally expensive for genome-wide datasets; gene score precision depends on the number of permutations.
VEGAS [23]	Takes into account the observed correlation between SNPs (LD) and simulates a specified number of statistics from which the resulting *p*-value is calculated.	Precision of gene scores depends on the number of simulations; computationally inefficient due to simulations.
Pegasus [24]	Pegasus leverages pathway-based information to prioritize weak signals in GWAS.	Performance heavily influenced by the quality and the relevance of pathway databases.
SKAT [15]	Employs mixed-model regression, considering covariates and genotypes for SNPs in a gene set to assess disease association.	May have limited power for small sample sizes and rare variant detection. Assume linear relationships between SNPs and the phenotype.

Table 2
A summary of the terminology and notation used in this article.

$\mathbb{S}$	a population of n individuals
$\mathbb{C}$	a set of j phenotype categories
$\mathbb{S}_c$	a subset of $\mathbb{S}$ containing individuals belonging to category c
$\gamma : \mathbb{S} -> \mathbb{C}$	labeling of the individuals category
$\mathbb{P}$	the set of m SNPs taken into account in the study
$loc : \mathbb{P} \mapsto \mathbb{N}^+$	the genomic location of a SNP
$\psi : \mathbb{S} \times \mathbb{P} \mapsto \{0, 1\}$	the state (genetic variation present or absent) of each SNP for each subject
$\mathbb{G}$	the set of regions that are investigated in the study
$\eta_c(g) \in [0, 1]$	the generalized allele spectrum of g with respect to phenotypic category c
$FC_{c_1,c_2}(g)$	the fold change of the generalized allele spectrum of region g with respect to two categories c_1 and c_2
$start : \mathbb{G} \mapsto \mathbb{N}^+$	the position of the first nucleotide in the region's genomic sequence.
$end : \mathbb{G} \mapsto \mathbb{N}^+$	the position of the last nucleotide in the region's genomic sequence.
$\rho(g \in \mathbb{G})$	the set of SNPs whose genomic location reside within the genomic location of the region G

Section 2.1 provides a formal description of the DiGAS method and a summary of the basic notions, as reported in Table 2. The methodology takes as input the coordinates of the genomic regions to be analyzed and the genotyping data (SNPs information) and the single-variant information regarding such regions. Subsequently, the pipeline involves the computation of the generalized allele spectrum, which is a measure related to the presence of SNPs in genomic regions for each phenotype condition analyzed in the study. Significant regions, that are differentially altered between two conditions, are identified based on the fold change of the generalized allele spectrum and the calculation of p-values using permutation tests and output by the pipeline.

Section 2.2 describes the data used and the preprocessing procedures applied.

Finally, Section 2.3 provides a description of the classification algorithms and evaluation metrics used to assess the performance of DiGAS.

2.1. DiGAS

Individuals with different phenotypic states can be categorized based on their conditions. For example, when studying a specific disease, we typically classify individuals into two groups: healthy and sick. However, it is also possible to consider more than two categories while ensuring that each individual belongs exclusively to one category.

In our model, the population of n individuals, referred to as subjects, is represented by the set $\mathbb{S} = \{s_1, s_2, \dots, s_n\}$, where s_i represents the ith individual. To categorize these individuals, we have a set of categories $\mathbb{C} = \{c_1, c_2, \dots, c_k\}$. We define a function $\gamma : \mathbb{S} -> \mathbb{C}$ to assign a category to each subject. A subset of $\mathbb{S}$ containing only the individuals belonging to category $c \in \mathbb{C}$ is denoted as $\mathbb{S}_c$.

For each individual, we examine the occurrence of single variations, i.e. single nucleotide polymorphisms (SNPs), in relation to a selected reference genome. We establish the function $loc : \mathbb{P} \to \mathbb{N}^+$ to determine the position of a SNP within the genome. We define $\mathbb{P} = \{p_1, p_2, \dots, p_m\}$ as the set of m SNPs that are being considered. It is important to note that in diploid genomes, where two alleles are present for each genomic locus, we do not differentiate between diploid variations at the same locus.

The function $\psi : \mathbb{S} \times \mathbb{P} \mapsto \{0, 1\}$ indicates the absence or presence of a SNP for an individual.

For instance, given an individual $s_i \in \mathbb{S}$ and a SNP $p_j \in \mathbb{P}$, $\psi(s_i, p_j)$ is 0 if no SNP is observed at $loc(p_j)$ in the genome of the individual s_i.

Experimental designs may necessitate the detection of SNPs throughout the entire genome or in specific regions such as genes, exons, or intergenic regions. The scope of SNP detection can be tailored based on the objectives of the study and the specific genomic regions of interest.

Consider the set of regions to investigate as $\mathbb{G} = \{g_1, g_2, \dots, g_l\}$, where each g_i represents a contiguous region of nucleotides defined by start and end coordinates with respect to the reference genome. We denote the subset of SNPs residing in the region g_i of the reference genome as $\rho(g_i \in \mathbb{G}) = \mathbb{P}_i \subseteq \mathbb{P}$. This subset $\mathbb{P}_i$ consists of SNPs where the genomic location $loc(p_j)$ satisfies the condition $start(g_i) \leq loc(p_j) \leq end(g_i)$ for each SNP $p_j \in \mathbb{P}_i$. In simpler terms, $\mathbb{P}_i$ includes SNPs located within the boundaries of the region g_i in the reference genome.

For a genomic region g belonging to the set $\mathbb{G}$, the overall allele spectrum of g in relation to the specified phenotype category c represents the ratio between the total count of SNPs observed in the region across all individuals within that category and the maximum possible

count of SNPs in that region for the same category. This can be defined as:

$$\eta_c(g) = \frac{\sum_{s_i \in \mathbb{S}_c} \sum_{p_j \in \rho(g)} \psi(s_i, p_j)}{|\rho(g)| \times |\mathbb{S}_c|} \qquad \in [0, 1]$$

with $\eta_c(g)$ is in the range $[0, 1]$ because $\psi(s_i, p_j)$ can be 0 or 1 and the summation cannot exceed $|\rho(g)| \times |\mathbb{S}_c|$. The value is 1 when all subjects belonging to the given category present all SNPs in the considered region.

Suppose that we aim to compare the overall allele spectrum for a specific genomic region g belonging to the set $\mathbb{G}$, between two phenotype sample categories c_1 and c_2. For such purpose, we define the fold change FC of a genomic region g with respect to the two categories c_1 and c_2 as:

$$FC_{c_1,c_2}(g) = |log(\frac{\eta_{c_1}(g) + 1}{\eta_{c_2}(g) + 1})| = |log(\eta_{c_1}(g) + 1) - log(\eta_{c_2}(g) + 1)|$$

where $\eta_{c_1}(g)$ is the overall allele spectrum of region g in the phenotype category c_1 and $\eta_{c_2}(g)$ is the overall allele spectrum of region g in the phenotype category c_2. Algorithm 1 presents the DiGAS pseudocode for the estimation of the overall allele spectrum of a region and the calculation of the allele spectrum fold change between two phenotype categories.

Algorithm 1 Procedure for computing FC values.

procedure ComputeFC($g, \mathbb{S}, \gamma, \rho, \psi, c_1, c_2$)
  $\mathbb{S}c_1 \leftarrow \{s \in \mathbb{S} : \gamma(s) = c_1\}$
  $\mathbb{S}c_2 \leftarrow \{s \in \mathbb{S} : \gamma(s) = c_2\}$
  $\eta_{c_1} \leftarrow \frac{\sum_{s_i \in \mathbb{S}_{c_1}} \sum_{p_j \in \rho(g)} \psi(s_i, p_j)}{|\rho(g)| \times |\mathbb{S}_{c_1}|}$
  $\eta_{c_2} \leftarrow \frac{\sum_{s_i \in \mathbb{S}_{c_2}} \sum_{p_j \in \rho(g)} \psi(s_i, p_j)}{|\rho(g)| \times |\mathbb{S}_{c_2}|}$
  return $|log(\eta_{c_1} + 1) - log(\eta_{c_2} + 1)|$
end procedure

Since our model allows us to compute the fold change of each region across each pair of phenotype categories, the selection of statistically significant regions is obtained by calculating an empirical p-value through a permutation test [32].

In Algorithm 2 we show the complete DiGAS pseudocode, including the steps needed to perform the permutation test. To achieve this, we initiate the process by randomly permuting the original category assignments of the subjects. This results in the creation of 1000 different random labelings of subject categories, denoted as $\{\gamma_0, \gamma_1, \ldots, \gamma_{1000}\}$. To determine the significance of the observed fold change in the real data, we calculate the proportion of random labelings where the fold change is equal to or greater than the observed value. This proportion represents the p-value of the region. A lower p-value indicates that the observed fold change is less likely to occur by random chance alone, suggesting that the region may have a significant association with the categories being studied.

More precisely, we modify the original category assignment γ to a new function γ_i, where the assignments in γ_i are a permutation of the assignments in γ. Thus, the total number of subjects assigned to each category, given two categories c_1 and c_2, is maintained from γ to γ_i.

Let $\mathbb{S}c_1$ and $\mathbb{S}c_2$ be the subsets obtained according to the category assignments in γ. To obtain γ_i, we iteratively modify γ for a total of $\frac{|\mathbb{S}c_1 \cup \mathbb{S}c_2|}{2}$ iterations. We refer to γ_i^t as the version of γ_i at iteration t, where γ_i^0 is an exact equal to γ. For each iteration $t > 0$, we select two subjects s_1 and s_2 such that $\gamma_i^0(s_1) \neq \gamma_i^0(s_2)$. We create γ_i^t by swapping $\frac{|\mathbb{S}c_1 \cup \mathbb{S}c_2|}{2}$ *times* the assignments of *the selected* s_1 and s_2, i.e., $\gamma_i^t(s_1) = \gamma_i^0(s_2)$, $\gamma_i^t(s_2) = \gamma_i^0(s_1)$, and $\gamma_i^t(s_i) = \gamma_i^0(s_i)$ for $s_i \in \mathbb{S} \setminus \{s_1, s_2\}$. The p-value of a region g is then determined by calculating the percentage of random labelings for which the fold change of the region equals or exceeds $FC_{c_1,c_2}(g)$. Regions that have a p-value less than 0.05 are considered relevant for discriminating between subjects who belong to category c_1 from subjects who belong to category c_2.

A flowchart summarizing the methodology is provided in Fig. 1.

Algorithm 2 Procedure for recognizing regions that differentiate between two phenotype categories using allele spectra.

procedure RecognizeRegions($\mathbb{G}, \mathbb{S}, \gamma, \rho, \psi, c_1, c_2$)
  $\hat{\mathbb{G}} \leftarrow \{\emptyset\}$
  $\mathbb{S}c_1 \leftarrow \{s \in \mathbb{S} : \gamma(s) = c_1\}$
  $\mathbb{S}c_2 \leftarrow \{s \in \mathbb{S} : \gamma(s) = c_2\}$
  $count \leftarrow 0$
  for each $g \in \mathbb{G}$ **do**
    $FC \leftarrow computeFC(g, \mathbb{S}, \gamma, \rho, \psi, c_1, c_2)$
    for $iter \leftarrow 1$ to 1000 **do**
      $\gamma' \leftarrow \gamma$ ▷ Gets a copy of γ
      for $i \leftarrow 1$ to $\frac{|\mathbb{S}c_1 \cup \mathbb{S}c_2|}{2}$ **do**
        $s_1 \leftarrow randomly\_select\_one(\mathbb{S}c_1)$
        $s_2 \leftarrow randomly\_select\_one(\mathbb{S}c_2)$
        $\gamma'(s_1) \leftarrow c_2$
        $\gamma'(s_2) \leftarrow c_1$
      end for
      $FC_{iter} \leftarrow computeFC(g, \mathbb{S}, \gamma', \rho, \psi, c_1, c_2)$
      if $FC_{iter} \geq FC$ **then**
        $count \leftarrow count + 1$
      end if
    end for
    if $count/1000 \leq 0.05$ **then**
      $\hat{\mathbb{G}} \leftarrow \hat{\mathbb{G}} \cup \{g\}$
    end if
  end for
  return $\hat{\mathbb{G}}$
end procedure

2.2. Test dataset

The data used in this manuscript was obtained from both The Alzheimer's Disease Neuroimaging Initiative (ADNI) project (http://adni.loni.usc.edu) and the Parkinson's Progression Markers Initiative (PPMI) project (http://www.ppmi-info.org, accessed November 19, 2019). The ADNI was launched in 2003 as a public–private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). The ADNI project includes various types of data, including MRI and PET images, genetics, cognitive tests, cerebrospinal fluid (CSF), and blood biomarkers for the study and prediction of the Alzheimer's disease. Specifically, our focus is on identifying genomic regions whose sets of SNPs collectively may contribute to the disease. Coordinates of the regions to take into account are provided by the GENCODE project[2] (*v36lift37*). In addition to the ADNI cohorts, we analyzed data from the PPMI project to further evaluate the versatility of our methodology. The PPMI project collects comprehensive data to study Parkinson's disease, including clinical assessments, imaging, and biospecimen data. In both the analysis we considered the complete set of ADNI cohorts, which includes ADNI1, ADNI2/GO, and ADNI3, as well as the PPMI cohort. The individuals in the ADNI cohorts are classified into three categories: affected (AD), not affected (CN), and mild cognitive impairment (MCI). The MCI category encompasses individuals who exhibit symptoms similar to those of Alzheimer's disease but do not exhibit a strong hallmark phenotype. In some cases, individuals with MCI may revert to normal conditions [33]. While the individuals in the PPMI cohort are classified into two categories: affected (PD) and not affected (CN).

[2] https://www.gencodegenes.org

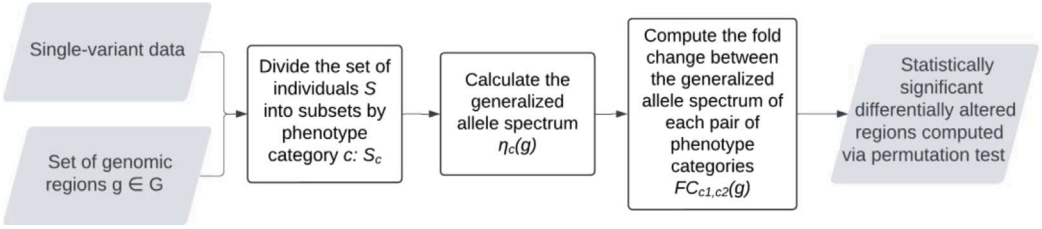

Fig. 1. Flowchart of the DiGAS methodology. The process starts with single-variant (SNP) data in PLINK format and a set of genomic regions G. SNPs are assigned to genomic regions $g \in G$ based on their location in the reference genome. The set of individuals S is divided into subsets S_c by phenotype category c. The generalized allele spectrum $\eta_c(g)$ is calculated for each region and phenotype categories, followed by computing the fold change between the generalized allele spectrum of each pair of phenotype categories. Finally, statistically significant differentially altered regions are identified via permutation test.

Table 3
Number of European subjects (divided by categories) used as input for each ADNI cohort. Total number of subjects, independently from their ancestry, is also reported.

	CN	MCI	AD	European subjects	Total subjects
ADNI1	197	339	168	704	757
ADNI2/GO	233	385	118	736	793
ADNI3	226	59	17	302	327

Table 4
Number of European subjects (divided by categories) used as input for the PPMI cohort. Total number of subjects, independently from their ancestry, is also reported.

	CN	PD	European subjects	Total subjects
PPMI	162	362	524	560

Table 5
Total number of SNPs for each cohort and number of SNPs filtered by Quality Control (QC) procedures.

	Original data	After QC
ADNI1	620.668	525.216
ADNI2/GO	730.525	591.481
ADNI3	759.993	303.150
PPMI	457.171	267.607

We filtered out all the individuals with no European ancestry. Statistics regarding the subjects extracted from ADNI and PPMI are reported in Tables 3 and 4 .

Quality control (QC) procedures were conducted on the data from each ADNI and PPMI cohort using PLINK 1.9 [34], which is a comprehensive toolset for whole-genome association analysis. These QC procedures involved filtering SNPs and subjects based on the following specific criteria: (i) Missing Data Filter ($geno > 0.2$): SNPs with a high proportion of missing data, where more than 20% of the data was missing, were excluded from the analysis. (ii) Individual Missingness Filter ($mind > 0.1$): SNPs were filtered based on individual missingness, where SNPs with more than 10% of individuals having missing genotype data were excluded.(iii) Minor Allele Frequency Filter ($MAF > 0.05$): SNPs with a minor allele frequency below 5% were removed. This filter helps to ensure that the analysis focuses on common genetic variations. (iv) Hardy-Weinberg Equilibrium Filter ($hwe > 1e{-}06$): SNPs showing significant deviations from the Hardy-Weinberg equilibrium were excluded. Hardy-Weinberg equilibrium represents the expected frequencies of genotypes in a population, and deviations from this equilibrium may indicate potential genotyping errors or other issues.

Table 5 provides information on the SNPs that were filtered out after applying these QC procedures. Regarding subjects, no individuals were filtered out based on QC measures. This means that all individuals in the ADNI and PPMI cohorts were retained for further analysis after the QC procedures.

2.3. Evaluation methodology

We used a set of classification algorithms, such as linear discriminant analysis (LDA) [35], support-vector machine (SVM) [36] (linear and polynomial), decision tree [37] and k-nearest neighbors (k-NN) [38] to evaluate the ability of the proposed methodology in selecting regions that are useful for distinguishing subject's categories. The goal of the classification is to build a model that, after a learning phase, correctly assigns a category to a given subject.

For this evaluation, we applied a 10-fold cross-validation [39]. This approach splits the original cohort into 10 subsets maintaining the initial proportions among the categories of subjects. Each subset is used in turn as the validation set, while the remaining nine subsets are combined to form the training set. The process is repeated ten times, ensuring that each subset is used exactly once as the validation data.

After the training phase, the resultant model is queried by using records belonging to the test set. A test set individual that is correctly recognized as belonging to a given category C by the model is considered a true positive (TP) for such a category. On the contrary, a false positive (FP) record is labeled as C by the model but, in reality, it does not belong to C. Similarly, true negatives (TN) are records that are correctly classified as non-C, and false negatives (FN) are records that are wrongly classified as not belonging to C.

Accuracy is defined as the fraction of records that are correctly classified with respect to the entire test set. The F1 score combines precision and recall statistics into a metric via harmonic mean. Precision informs about the fraction of records that are correctly classified as belonging to C with respect to the total number of records that are classified as C by the model. Recall gives the fraction of records belonging to C that are correctly classified with respect to the total size of C.

All the given metrics are in the range of $[0, 1]$ such that the higher the value, the better the performance of the given model is. Moreover, for binary classification, precision and recall are related to the given category that is taken into account. On the contrary, the value of accuracy is the same independently for the investigated category.

3. Results

DiGAS methodology was evaluated by assessing its accuracy with respect to SKAT [15], in identifying groups of SNPs effective in classifying individuals. Given an ADNI cohort (see Section 2.2 for details regarding the composition of the ADNI dataset), we employed a 10-fold cross-validation approach splitting the dataset into 10 equal parts, using 9 parts for training and 1 part for validation and rotating this process so that each part is used as the validation set once. This ensures that each subset of the data is used for both training and validation, providing a robust evaluation of the model's performance.

This means that, if the ADNI1 cohort has $197/704 = 28\%$ CN subjects, 48% MCI and 23% AD, such percentages are preserved in both the training and the validation sets.

SNP sets are the features of our classification model. Thus, the goal is to recognize the SNP sets which make a distinction between the

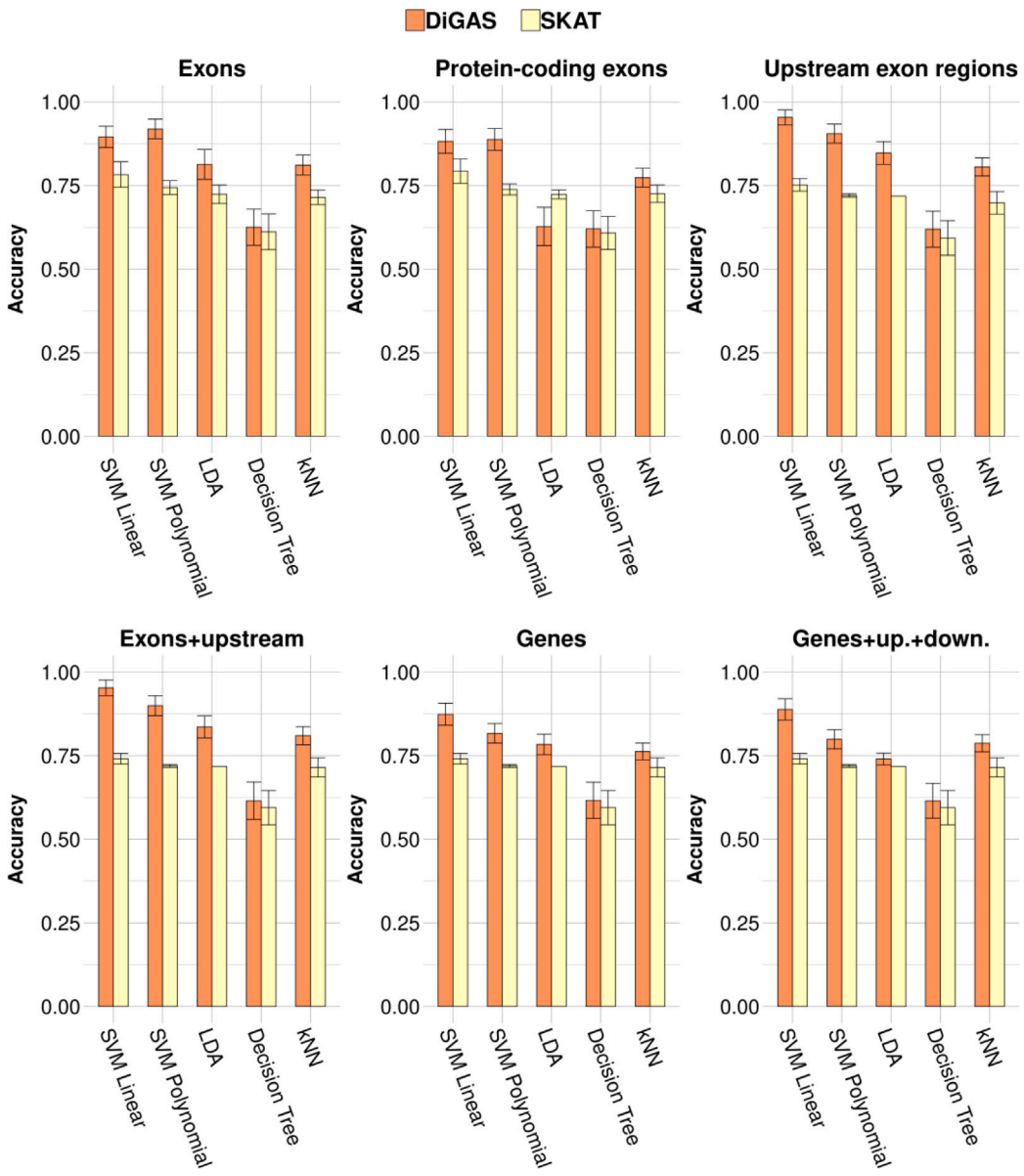

Fig. 2. Accuracy metrics on ADNI1 using 10-fold cross-validation for each evaluated classification algorithm and each genomic region.

categories and to do so, we grouped SNPs by the following genomic regions:

- Exons: each exon is considered a distinct region, not linked to the exons of the same gene. Exons may belong to any type of gene, protein-coding or not.
- Protein-coding exons: namely exons that belong to genes known to code for proteins.
- Upstream exon regions: for each exon we extracted 5k nucleotides preceding the exon, excluding the exon itself.
- Exons+upstream: for each exon, we included the exon plus the upstream 5K nucleotides region.
- Genes: the complete genomic sequence of each gene, including exons and introns.
- Genes+upstream+downstream: we extract the upstream and the downstream regions, for 20 Kb each, along with the gene sequence, as used in [15].

Coordinates of such genomic elements were extracted from public databases described in Section 2.2 and the belonging of a SNP to a given region is calculated via the *loc* function described in Section 2.1. All the experiments were performed over the GrCh38 version of the human genome. Since ADNI1 is originally defined over previous versions of the human genome, we used the tool UCSC LiftOver [40,41] to convert such coordinates into coordinates over the GrCh38 genome.

We applied the methodology described in Section 2.1 to identify significant regions, considering a p-value cut-off of 0.05 (evaluated by fold-change). In this process, the three categories, CN, AD and MCI, were evaluated separately. Then, we merged the regions that resulted significant for AD and MCI into a single set of regions and for this reason, in what follows, ill subjects are also referred to as the joint category AD/MCI. For each cross-validation, the resultant performance metrics were calculated by running 1,000 iterations, and by computing the mean and the standard deviation of the results.

Fig. 2 shows the accuracy values of DiGAS and SKAT on the ADNI1 cohort varying the genomic regions and the type of classifier. For each type of region considered, DiGAS always outperforms SKAT in particular using SVM classifiers. Regardless of the genomic region, decision tree classifiers yield the lowest accuracy for both methods. SKAT reaches a maximum accuracy of 0.79 when exons or protein-coding exons are used as the basis for training an SVM linear classifier,

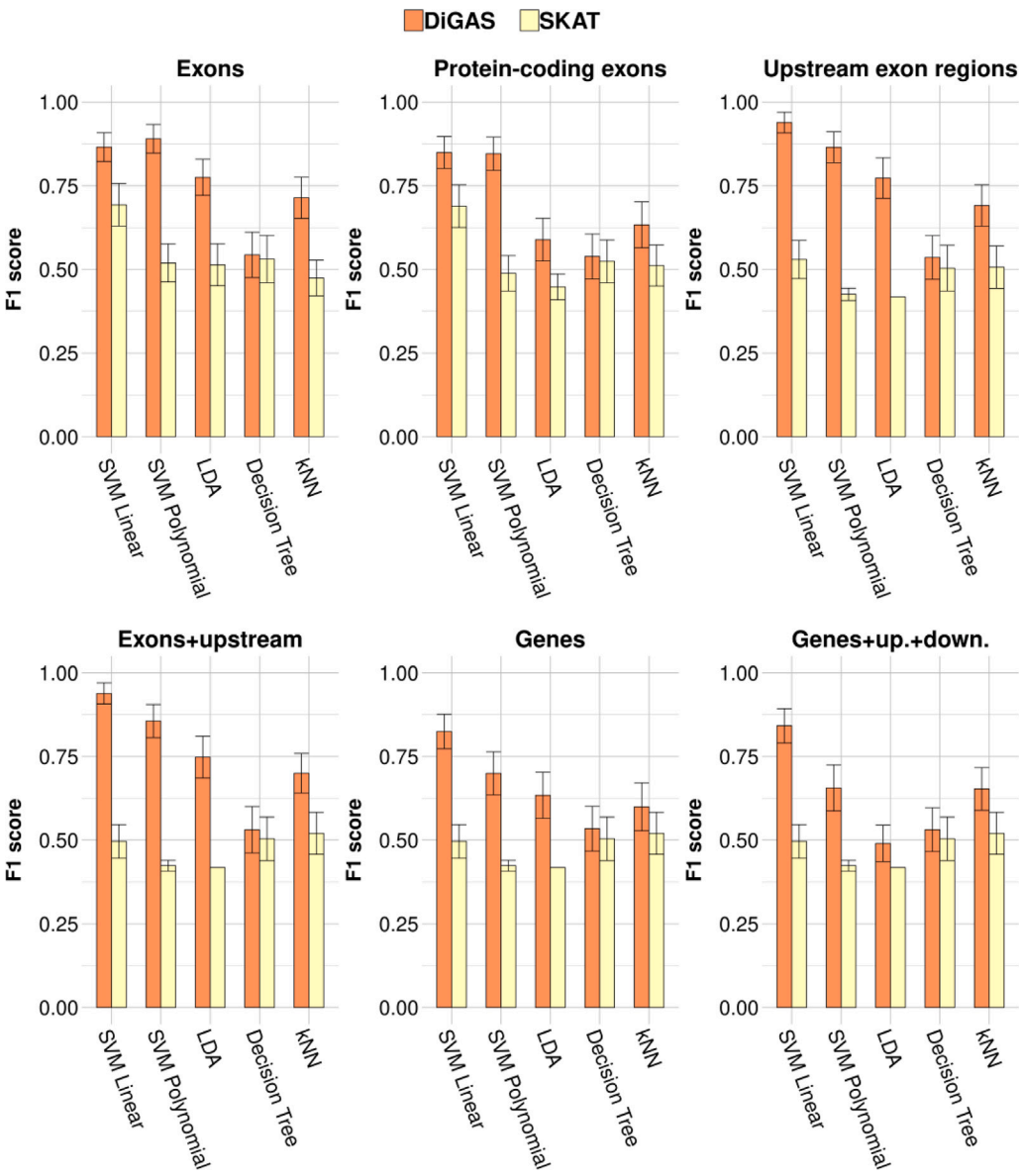

Fig. 3. F1 score metrics on ADNI1 using 10-fold cross-validation for each evaluated classification algorithm and each genomic region.

but in general SKAT accuracy is mostly below 0.75. On the contrary, DiGAS is able to break the barrier of 0.75 in multiple configurations. The best accuracy value of 0.95 is obtained when upstream exon regions are taken into account alone or in combination with exons and by using an SVM classifier.

Similar results are shown for the F1 score for the ADNI1 cohort in Fig. 3. SKAT reaches a maximum F1 score of 0.79 via exons or protein-coding exons, while DiGAS obtains up to a 0.94 of F1 score on both upstream exon regions and exon+upstream.

Figs. 4 and 5 report accuracy and F1 score values for the ADNI2 cohort, reflecting performance trends similar to ADNI1. Maximum values of accuracy are 0.93 (exons+upstream and upstream exon regions) and 0.79 (exons) for DiGAS and SKAT, respectively. Maximum F1 scores are 0.92 (exons+upstream and upstream exon regions) and 0.73 (exons) for DiGAS and SKAT, respectively.

Figs. 6 and 7 show results obtained on the ADNI3 cohort. Accuracy values follow similar trends obtained by testing the methodologies on ADNI1 and ADNI2 however, DiGAS and SKAT reduce their performance considering the F1 score. The difference with previous cohorts is due to the limited number of AD/MCI subjects included in the dataset. ADNI3 cohort is an ongoing project for which fewer ill subjects are yet reported. F1 scores for the AD/MCI group suffer such a lack of data that does not affect accuracy because such a measure takes into account both CN and AD/MCI groups. However, it has to be noticed that DiGAS is still able to reach an F1 score of 0.93 when exons+upstream regions are combined with the kNN classifier, and exon regions produce a maximum value of 0.92 when SVM linear and kNN classifiers are employed. Moreover, DiGAS crosses the barrier of 0.70 in several configurations. On the contrary, SKAT reaches an F1 score greater the 0.70 only in five configurations, being the best one equal to 0.72 by combining exon regions with the SVM linear classifier or upstream exon regions with the kNN linear classifier. These results demonstrate DiGAS's robustness with limited data. In general, the SVM linear classifier is the best choice to work with the DiGAS methodology, but the kNN approach can be taken into account in the presence of a dataset with a category containing a limited number of subjects.

We note that exons, and in particular not only protein-coding exons, combined with upstream regions yield the best classification results. Alzheimer's disease is a complex disease which involves many genes and, presumably, their regulatory elements [42,43] and such elements are often placed in upstream gene regions. However, our analysis

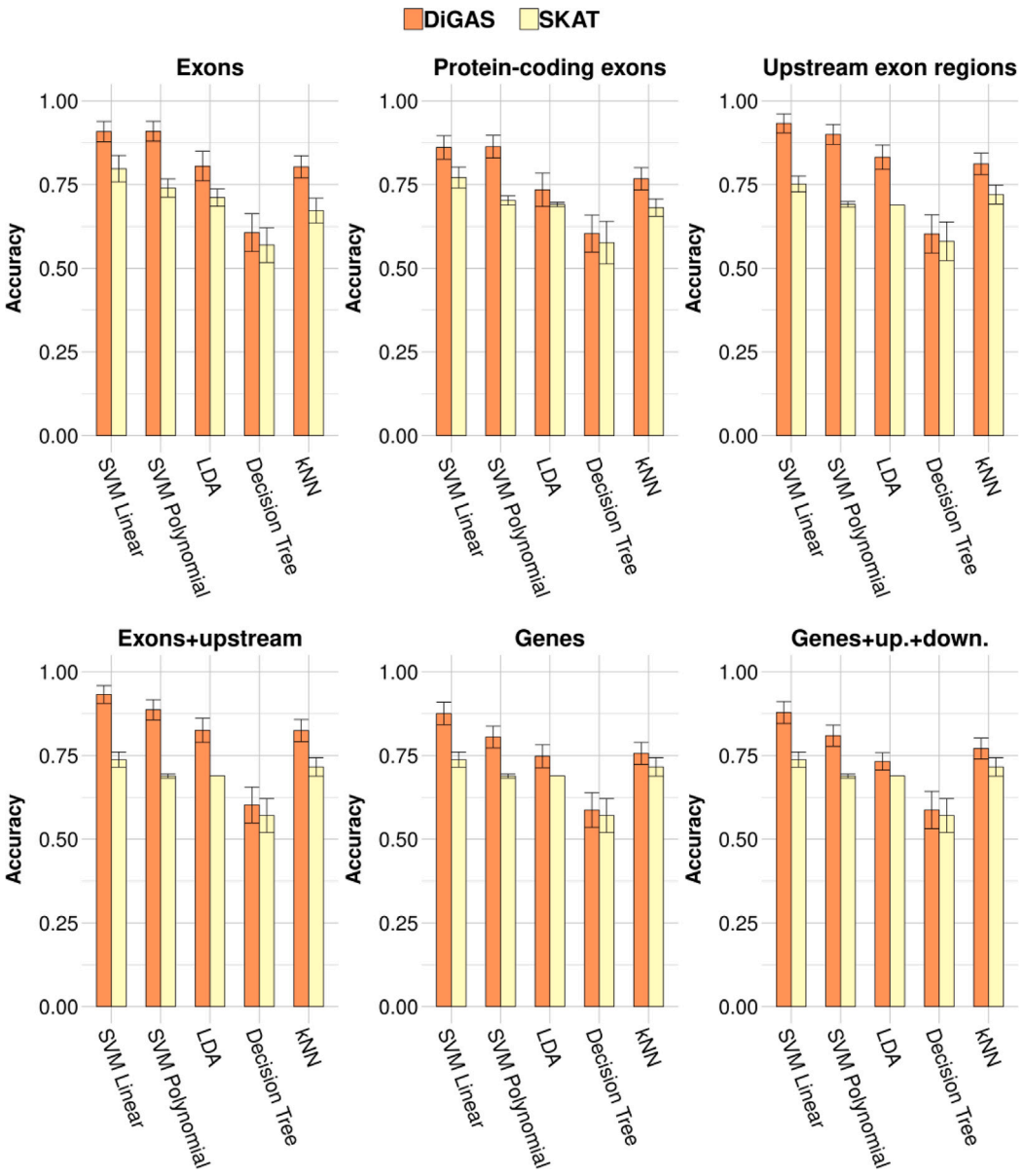

Fig. 4. Accuracy metrics on ADNI2 using 10-fold cross-validation for each evaluated classification algorithm and each genomic region.

shows that regulatory regions of genes are important as well as upstream exon regions. It is known that too much information may reduce classifiers' performance, especially when such overabundant data does not relate to the recognition problem that is taken into account. Low performance on genes and their combined regions suggests the significance of upstream exon regions and inter-/intra-genic regulatory elements in Alzheimer's disease. Upstream regions alone produce results comparable to their combination with exons, indicating overlapping information. Pure exon regions are outperformed by their combination with upstream regions.

To further evaluate the versatility and effectiveness of DiGAS, we conducted additional experiments using the Parkinson's Progression Markers Initiative (PPMI) dataset. Fig. 8 shows the F1 score values of DiGAS and SKAT on the PPMI cohort across the different genomic regions and classifier types. DiGAS consistently outperforms SKAT, particularly when using SVM classifiers. The highest F1 score achieved by DiGAS is 0.83 using genes+upstream+downstream regions, whereas SKAT's highest F1 score is 0.80 using the same genomic regions. Unlike the results observed with the ADNI cohorts, in the PPMI dataset, the F1 score tends to increase as the size of the genomic region increases, from exons to entire genes. This improvement is particularly noticeable when including the upstream and downstream regions of the genes.

We also evaluated the computational resources required for running DiGAS, including memory and processing power. The results are shown in Fig. 9. DiGAS demonstrates significantly better performance compared to SKAT in terms of computational time and memory usage, especially when dealing with distinct and smaller genomic regions. As the regions decrease in number and the number of SNPs per region increases, SKAT performance becomes comparable to DiGAS, particularly regarding execution time. Overall, these results demonstrate that DiGAS is superior to SKAT across different phenotypic conditions and datasets, and that the SVM classifier consistently provides the best results.

Finally, we performed 10 leave-one-out and k-folding tests to identify core genes significant in all DiGAS iterations. For each iteration,

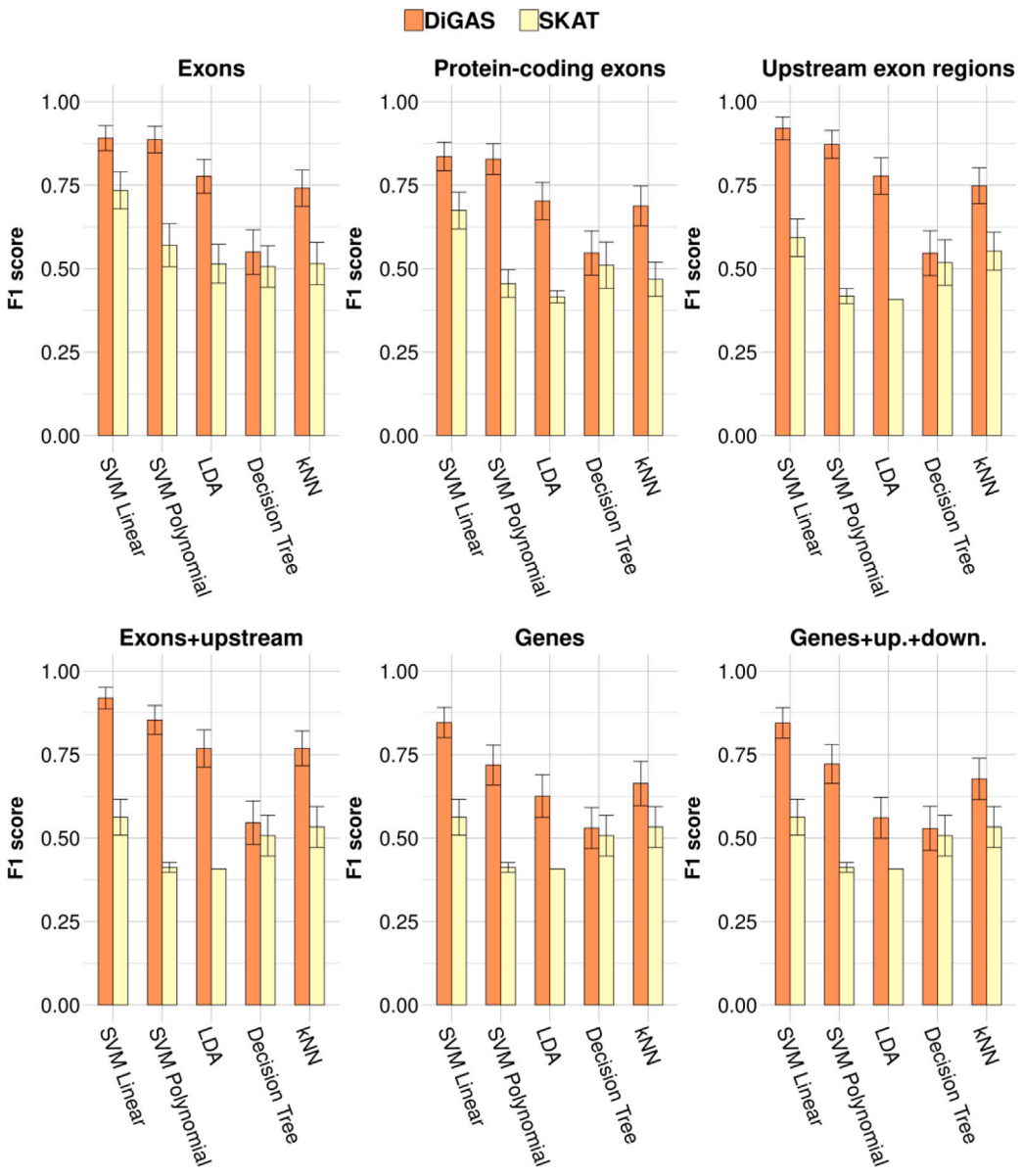

Fig. 5. F1 score metrics on ADNI2 using 10-fold cross-validation for each evaluated classification algorithm and each genomic region.

in the leave-one-out test we randomly select a subject, while in the k-folding test, we randomly extract 90% of subjects maintaining percentages among phenotype categories. Fig. 10 shows results on ADNI1 cohort where red points are the core genes and blue points are genes not significant in at least one iteration. The results show that as the *p*-value decreases, DiGAS obtains a set of genes that does not vary if subjects are removed from the input dataset.

Fig. 11 shows that DiGAS maintains superior performance with respect to SKAT both in accuracy and F1 score even using only k-folding core genes to classify Alzheimer's disease subjects.

4. Conclusion

We propose DiGAS, a novel methodology for the identification of SNP sets associated with a specific phenotype condition that comprehensively and simultaneously analyzes the entire set of SNPs within genomic regions. Taking advantage of the introduction of the generalized allele spectrum descriptor and identifying sets of features based on allele frequency differences, DiGAS enhances the accuracy of genetic signal attribution to specific genomic regions, overcoming limitations inherent in SNP-level analyses commonly employed by other methods, which assigns a significance score to each individual SNP before grouping them into SNP sets. Tests conducted on well-established datasets related to Alzheimer's disease and Parkinson's disease, respectively collected from ADNI and PPMI, show that the tool consistently outperforms SKAT in computational efficiency and in identifying predictive genomic features for individuals classification. Moreover, DiGAS does not make annotations on genes or patients and it is designed to be independent of ethnic background, ensuring that the analysis is not influenced by the ethnic composition of the datasets. The tool is highly distributable and designed to be easily integrated within bioinformatics pipelines. Although DiGAS demonstrates promising performance in the identification of genomic regions associated with a specific phenotype, several limitations and challenges need to be considered. As competitor methods, DiGAS accuracy is heavily dependent on the quality and quantity of the input genomic data and incomplete or low-quality datasets can affect the reliability of the results. Additionally, the identification of significant SNP-sets does not provide direct biological insights and further functional studies are required to interpret the biological relevance of the findings.

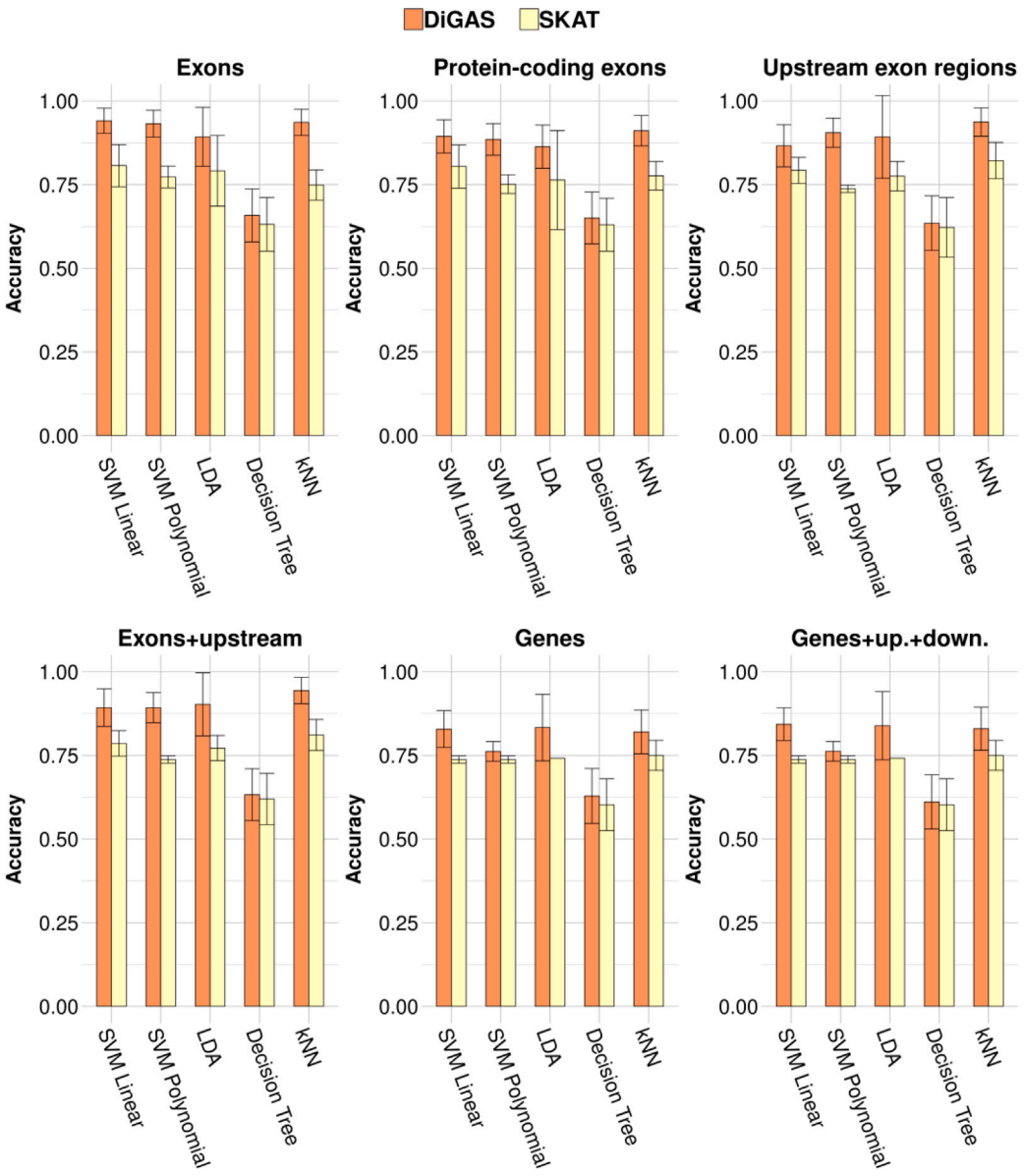

Fig. 6. Accuracy metrics on ADNI3 using 10-fold cross-validation for each evaluated classification algorithm and each genomic region.

CRediT authorship contribution statement

Antonino Aparo: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology. **Vincenzo Bonnici:** Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Simone Avesani:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology. **Luciano Cascione:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Conceptualization. **Rosalba Giugno:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work has been partially supported by #NEXTGENERATIONEU and the Italian Ministry of University and Research, National Recovery and Resilience Plan (PNRR), projects MNESYS (PE0000006), HEAL ITALIA (PE0000019), National Biodiversity Future Center – NBFC (CN_00000033), and the "PREPARE" project (n. F/310130/05/X56 - CUP: B39J23001730005) - D.M. MiSE 31/12/2021.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI), United States (National Institutes of Health Grant U01 AG024904) and DOD ADNI, United States (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, United States, the National Institute of Biomedical Imaging and Bioengineering, United States, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La

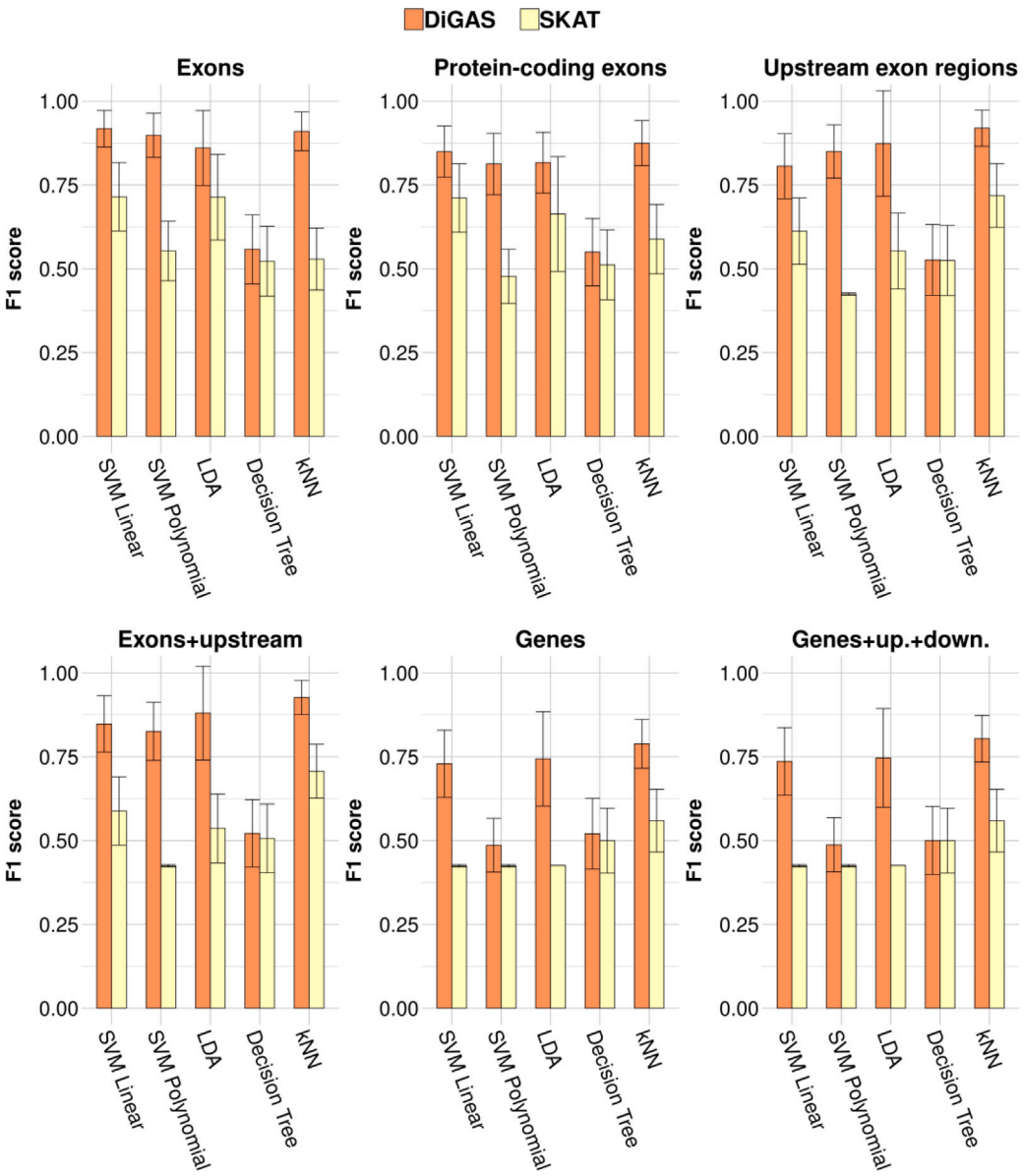

Fig. 7. F1 score metrics on ADNI3 using 10-fold cross-validation for each evaluated classification algorithm and each genomic region.

Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health, United States (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Data used in the preparation of this article were obtained on November 19, 2019 from the Parkinson's Progression Markers Initiative (PPMI) database (https://www.ppmi-info.org/access-data-specimens/download-data), RRID:SCR_006431. The data were obtained from PPMI upon request after approval by the PPMI Data Access Committee. For up-to-date information on the study, visit http://www.ppmi-info.org. PPMI – a public–private partnership – is funded by the Michael J. Fox Foundation for Parkinson's Research, United States and funding partners, including 4D Pharma, Abbvie, AcureX, Allergan, Amathus Therapeutics, Aligning Science Across Parkinson's, AskBio, Avid Radiopharmaceuticals, BIAL, BioArctic, Biogen, Biohaven, BioLegend, BlueRock Therapeutics, Bristol-Myers Squibb, Calico Labs, Capsida Biotherapeutics, Celgene, Cerevel Therapeutics, Coave Therapeutics, DaCapo Brainscience, Denali, Edmond J. Safra Foundation, Eli Lilly, Gain Therapeutics, GE HealthCare, Genentech, GSK, Golub Capital, Handl Therapeutics, Insitro, Jazz Pharmaceuticals, Johnson & Johnson Innovative Medicine, Lundbeck, Merck, Meso Scale Discovery, Mission Therapeutics, Neurocrine Biosciences, Neuron23, Neuropore, Pfizer, Piramal, Prevail Therapeutics, Roche, Sanofi, Servier, Sun Pharma Advanced Research Company, Takeda, Teva, UCB, Vanqua Bio, Verily, Voyager Therapeutics, the Weston Family Foundation and Yumanity Therapeutics.

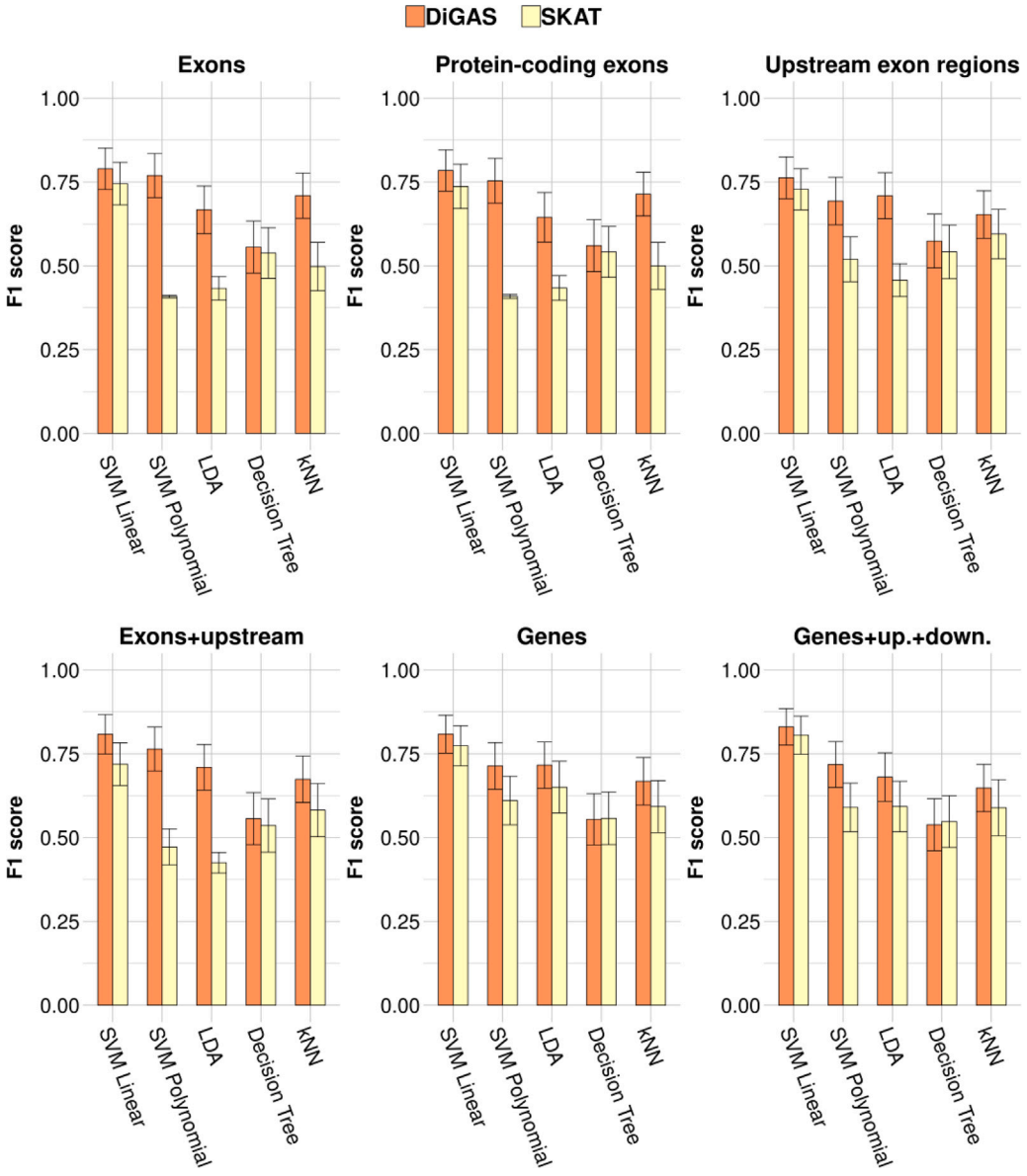

Fig. 8. F1 score metrics on PPMI using 10-fold cross-validation for each evaluated classification algorithm and each genomic region.

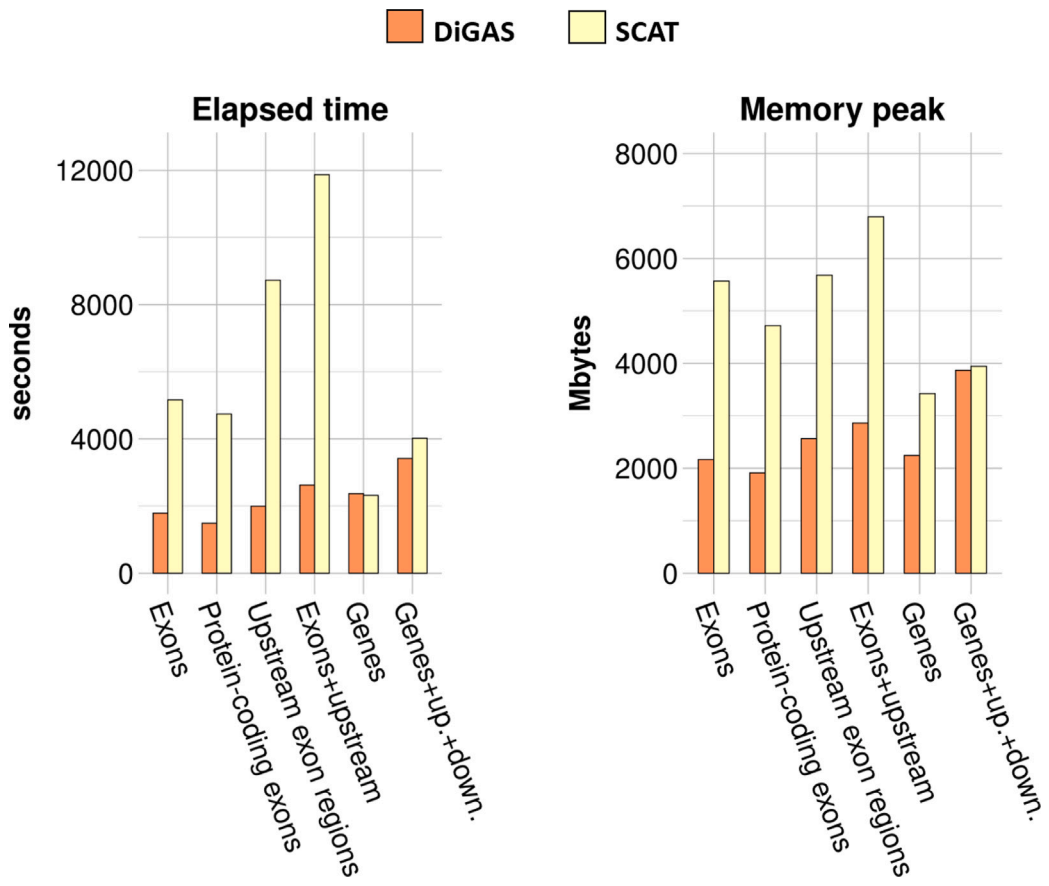

Fig. 9. Comparison of computational resources required by DiGAS and SKAT for different genomic regions in the PPMI dataset. The left plot represents elapsed time, and the right plot represents memory usage for both methods across various genomic region sizes.

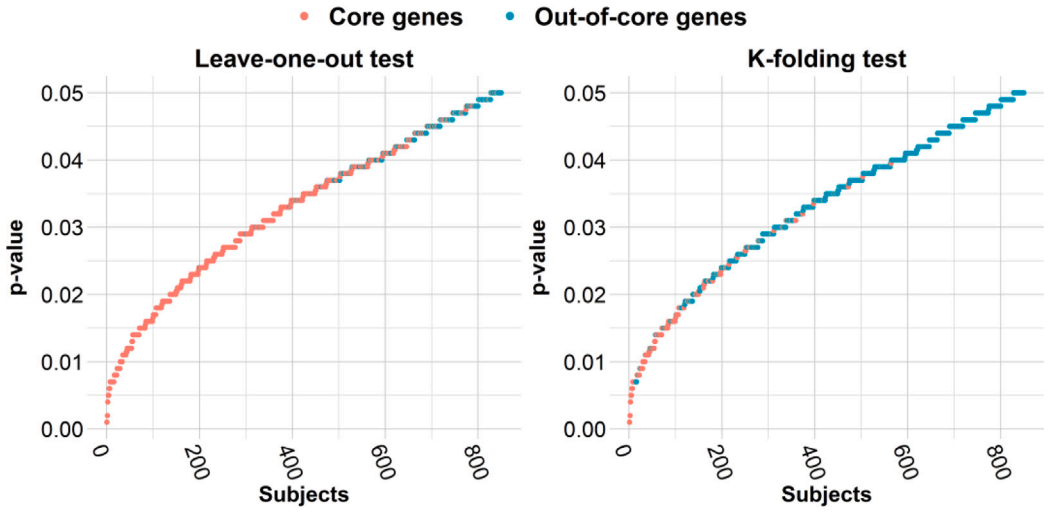

Fig. 10. Leave-one-out and k-folding test results on ADNI1 cohort. Each point refers to a gene and its *p*-value. Genes are ranked on the *x*-axis according to their *p*-value (y-axis). Red genes, also called core genes, are significant in all tests, otherwise, they are colored as blue.

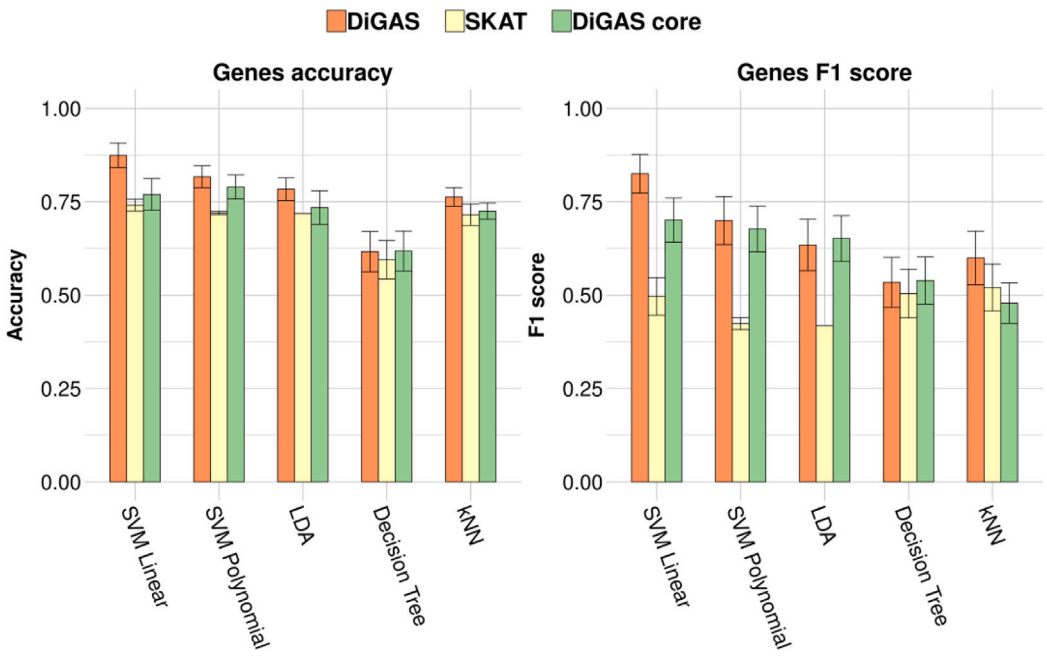

Fig. 11. Accuracy and F1 score metrics on ADNI1 using 10-fold cross-validation for each classification algorithm, and by using genes as the genomic region.

References

[1] D.A. Al-Koofee, S.M. Mubarak, Genetic polymorphisms, Recent Top. Genet. Polymorph. (2019) 1–10.

[2] P. Wainschtein, D. Jain, Z. Zheng, L.A. Cupples, A.H. Shadyab, B. McKnight, B.M. Shoemaker, B.D. Mitchell, et al., Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data, Nature Genet. 54 (3) (2022) 263–273.

[3] D. Kinane, T. Hart, Genes and gene polymorphisms associated with periodontal disease, Crit. Rev. Oral Biol. Med. 14 (6) (2003) 430–449.

[4] A. Chakravarti, ... To a future of genetic medicine, Nature 409 (6822) (2001) 822–823.

[5] T. Kanekiyo, H. Xu, G. Bu, ApoE and Aβ in Alzheimer's disease: accidental encounters or partners? Neuron 81 (4) (2014) 740–754.

[6] J. Poirier, P. Bertrand, S. Kogan, S. Gauthier, J. Davignon, D. Bouthillier, Apolipoprotein E polymorphism and Alzheimer's disease, Lancet 342 (8873) (1993) 697–699.

[7] P.M. Visscher, N.R. Wray, Q. Zhang, P. Sklar, M.I. McCarthy, M.A. Brown, J. Yang, 10 Years of GWAS discovery: biology, function, and translation, Am. J. Hum. Genet. 101 (1) (2017) 5–22.

[8] E. Uffelmann, Q.Q. Huang, N.S. Munung, J. de Vries, Y. Okada, A.R. Martin, H.C. Martin, T. Lappalainen, D. Posthuma, Genome-wide association studies, Nature Rev. Methods Primers 1 (1) (2021) 1–21.

[9] V. Tam, N. Patel, M. Turcotte, Y. Bossé, G. Paré, D. Meyre, Benefits and limitations of genome-wide association studies, Nature Rev. Genet. 20 (8) (2019) 467–484.

[10] X. Wang, N.J. Morris, D.J. Schaid, R.C. Elston, Power of single-vs. multi-marker tests of association, Genet. Epidemiol. 36 (5) (2012) 480–487.

[11] J. Zhou, O.G. Troyanskaya, Predicting effects of noncoding variants with deep learning–based sequence model, Nature methods 12 (10) (2015) 931–934.

[12] D. Quang, X. Xie, Danq: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences, Nucleic Acids Res. 44 (11) (2016) e107.

[13] P.C. Phillips, Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems, Nature Rev. Genet. 9 (11) (2008) 855–867.

[14] E. Kuzmin, B. VanderSluis, W. Wang, G. Tan, R. Deshpande, Y. Chen, M. Usaj, A. Balint, M. Mattiazzi Usaj, J. Van Leeuwen, et al., Systematic analysis of complex genetic interactions, Science 360 (6386) (2018) eaao1729.

[15] M.C. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, X. Lin, Rare-variant association testing for sequencing data with the sequence kernel association test, Am. J. Hum. Genet. 89 (1) (2011) 82–93.

[16] H.I. Avi-Itzhak, X. Su, F.M. De La Vega, Selection of minimum subsets of single nucleotide polymorphisms to capture haplotype block diversity, in: Biocomputing 2003, World Scientific, 2002, pp. 466–477.

[17] A. Torkamani, E.J. Topol, N.J. Schork, Pathway analysis of seven common diseases assessed by genome-wide association, Genomics 92 (5) (2008) 265–272.

[18] Y. Hu, L. Deng, J. Zhang, X. Fang, P. Mei, X. Cao, J. Lin, Y. Wei, X. Zhang, R. Xu, A pooling genome-wide association study combining a pathway analysis for typical sporadic parkinson's disease in the han population of Chinese mainland, Mol. Neurobiol. 53 (2016) 4302–4318.

[19] H. Schwender, I. Ruczinski, K. Ickstadt, Testing SNPs and sets of SNPs for importance in association studies, Biostatistics 12 (1) (2011) 18–32.

[20] K. Wang, M. Li, M. Bucan, Pathway-based approaches for analysis of genomewide association studies, Am. J. Hum. Genet. 81 (6) (2007) 1278–1283.

[21] K.-S. Wang, X. Liu, Q. Zhang, N. Aragam, Y. Pan, Parent-of-origin effects of FAS and PDLIM1 in attention-deficit/hyperactivity disorder, J. Psychiatry Neurosci. 37 (1) (2012) 46–52.

[22] D.H. Ballard, J. Cho, H. Zhao, Comparisons of multi-marker association methods to detect association between a candidate region and disease, Genet. Epidemiol.: Off. Publ. Int. Genet. Epidemiol. Soc. 34 (3) (2010) 201–212.

[23] J.Z. Liu, A.F. Mcrae, D.R. Nyholt, S.E. Medland, N.R. Wray, K.M. Brown, N.K. Hayward, G.W. Montgomery, P.M. Visscher, N.G. Martin, et al., A versatile gene-based test for genome-wide association studies, Am. J. Hum. Genet. 87 (1) (2010) 139–145.

[24] P. Nakka, B.J. Raphael, S. Ramachandran, Gene and network analysis of common variants reveals novel associations in multiple complex diseases, Genetics 204 (2) (2016) 783–798.

[25] A. Abeliovich, A.D. Gitler, Defects in trafficking bridge Parkinson's disease pathology and genetics, Nature 539 (7628) (2016) 207–216.

[26] J.P. Taylor, R.H. Brown, D.W. Cleveland, Decoding ALS: from genes to mechanism, Nature 539 (7628) (2016) 197–206.

[27] P. Scheltens, B. De Strooper, M. Kivipelto, H. Holstege, G. Chételat, C.E. Teunissen, J. Cummings, W.M. van der Flier, Alzheimer's disease, Lancet 397 (10284) (2021) 1577–1590.

[28] M.D. Ritchie, K. Van Steen, The search for gene-gene interactions in genome-wide association studies: challenges in abundance of methods, practical considerations, and biological interpretation, Ann. Transl. Med. 6 (8) (2018).

[29] X. Wan, C. Yang, Q. Yang, H. Xue, N.L. Tang, W. Yu, Predictive rule inference for epistatic interaction detection in genome-wide association studies, Bioinformatics (ISSN: 1367-4803) 26 (1) (2009) 30–37.

[30] Alzheimer's Disease Neuroimaging Initiative, Alzheimer's Disease Neuroimaging Initiative (ADNI), 2003, https://adni.loni.usc.edu/. The ADNI Project.

[31] Parkinson's Progression Markers Initiative, Parkinson's progression markers initiative (PPMI), 2011, https://www.ppmi-info.org/. The Michael J. Fox Foundation for Parkinson's Research.

[32] P. Good, Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses, Springer Science & Business Media, 2013.

[33] T.D. Koepsell, S.E. Monsell, Reversion from mild cognitive impairment to normal or near-normal cognition: risk factors and prognosis, Neurology 79 (15) (2012) 1591–1598.

[34] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M.A. Ferreira, D. Bender, J. Maller, P. Sklar, P.I. De Bakker, M.J. Daly, et al., PLINK: a tool set for whole-genome association and population-based linkage analyses, Am. J. Hum. Genet. 81 (3) (2007) 559–575.

[35] S. Balakrishnama, A. Ganapathiraju, Linear discriminant analysis-a brief tutorial, Inst. Signal Inf. Process. 18 (1998) (1998) 1–8.
[36] W.S. Noble, What is a support vector machine? Nature Biotechnol. 24 (12) (2006) 1565–1567.
[37] J.R. Quinlan, Induction of decision trees, Mach. Learn. 1 (1) (1986) 81–106.
[38] X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G.J. McLachlan, A. Ng, B. Liu, S.Y. Philip, et al., Top 10 algorithms in data mining, Knowl. Inf. Syst. 14 (1) (2008) 1–37.
[39] R. Kohavi, et al., A study of cross-validation and bootstrap for accuracy estimation and model selection, in: Ijcai, vol. 14, no. 2, Montreal, Canada, 1995, pp. 1137–1145.
[40] B.T. Lee, G.P. Barber, A. Benet-Pagès, J. Casper, H. Clawson, M. Diekhans, C. Fischer, J.N. Gonzalez, A.S. Hinrichs, C.M. Lee, et al., The UCSC genome browser database: 2022 update, Nucleic Acids Res. 50 (D1) (2022) D1115–D1122.
[41] P.-L. Luu, P.-T. Ong, T.-P. Dinh, S.J. Clark, Benchmark study comparing liftover tools for genome conversion of epigenome sequencing data, NAR Genom. Bioinform. 2 (3) (2020) lqaa054.
[42] G. Novikova, M. Kapoor, J. Tcw, E.M. Abud, A.G. Efthymiou, S.X. Chen, H. Cheng, J.F. Fullard, J. Bendl, Y. Liu, et al., Integration of Alzheimer's disease genetics and myeloid genomics identifies disease risk regulatory elements and genes, Nature Commun. 12 (1) (2021) 1–14.
[43] J.W. Touchman, A. Dehejia, O. Chiba-Falek, D.E. Cabin, J.R. Schwartz, B.M. Orrison, M.H. Polymeropoulos, R.L. Nussbaum, Human and mouse α-synuclein genes: comparative genomic sequence analysis and identification of a novel gene regulatory element, Genome Res. 11 (1) (2001) 78–86.