UNIVERSITÀ
DI PAVIA

PhD in Psychology, Neuroscience and Data Science
Department of Brain and Behavioral Sciences

Investigating ALS genetic epidemiology complexity through a multi-analytical approach and introducing GenUInE, a tool for genomic signals prioritization

---

Academic year 2023-2024
Cycle XXXVII

Coordinatore
Prof. Elena Cavallini

Doctoral candidate
Dr. Alberto Brusati

Tutor
Prof. Davide Gentilini

Co-tutor
Prof. Nicola Ticozzi

Dedicated to Lorenzina,
for her example of determination and tenacity.

Abstract

Amyotrophic Lateral Sclerosis (ALS) is a complex neurodegenerative disease characterized by motor neuron degeneration in both familial (fALS) and sporadic (sALS) forms. Despite the great advancements in identifying genetic factors, understanding the full genetic landscape of complex diseases such as ALS remains challenging, due to the limited power of the studies or intrinsic constraints of single analytical methods. To address this point, we developed GenUInE, a multi-analysis aggregator tool designed to integrate results from various genomic analyses into a final unified matrix, enabling the identification of genetic hotspots. GenUInE uses a probability-based model to prioritize genomic windows associated with disease traits by analyzing diverse input sources such as homozygosity mapping, IBD segments, epivariations, and rare variants. The tool computes combined probabilities and summation values for each window, additionally providing a score for prioritizing and weighting genomic regions. Applied to ALS, GenUInE successfully highlighted previously ALS-associated known genes (*NIPA1*) and identified novel genetic signatures linked to neurodegenerative pathways. Our work provides a novel framework for exploring genetics in complex diseases, providing a different method aimed at identifying new therapeutic targets.

Introduction

Background: ALS disease, from history to epidemiology

Amyotrophic Lateral Sclerosis (ALS) is a progressive neurodegenerative disorder that primarily affects motor neurons in the brain and spinal cord. The first clinical descriptions of ALS can be traced back to the mid-19th century. Jean-Martin Charcot is often credited with formally identifying the disease in 1869, although various neurologists previously described the fundaments of the disease [1]. In 1824, Charles Bell published an important step forward, introducing the concept that diseases could be exclusively motor-dependent. Specifically, Bell was the first to differentiate between two types of roots within the spinal cord: the anterior roots, responsible for motor functions related to movement, and the posterior roots with a sensory function. In 1848, François Aran hypothesized a neurogenic cause for a syndrome leading to progressive muscle weakness. Charcot, however, was the first to identify ALS as an independent disease and emphasize its neurological basis, specifically, linking the degeneration of motor neurons in the brain and spinal cord to the onset of the disease.

During the first half of the 20th century, ALS gained popularity due to high-profile cases such as the baseball player Lou Gehrig. His death shook public opinion so much that ALS was renamed Lou Gehrig's disease. However, all this increased the determination to find a cure, and consequently, efforts to better understand its molecular basis intensified. We now know that the progressive degeneration of both the first and second motor neurons characterizes ALS [2]. The first or upper motor neuron originates in the brain's motor cortex and travels down to the brainstem or spinal cord, particularly into the body of the second or lower motor neuron. Its primary function is to transmit inhibitory or excitatory signals that control voluntary movements. The second or lower motor neuron projects from the spinal cord and brainstem to innervate muscles and glands, directly causing skeletal muscle contraction [3]. Degeneration of upper motor neurons led to several symptoms, such as spasticity (an increase in muscle tone that leads to stiffness and tightness in muscles), hyperreflexia (an exaggerated reflex response), lack of coordination, slow movements, Dysarthria (difficulty in speech), and dysphagia (difficulty in swallowing).

Loss of lower motor neurons affects the direct connection to skeletal muscles and leads to muscle weakness starting in the hands or feet, muscle atrophy due to loss of innervation, fasciculations (involuntary muscle twitches), muscle cramps, and diminished reflexes. Altogether these symptoms cause complete body paralysis and death of the patients for respiratory arrest (Figure 1a).

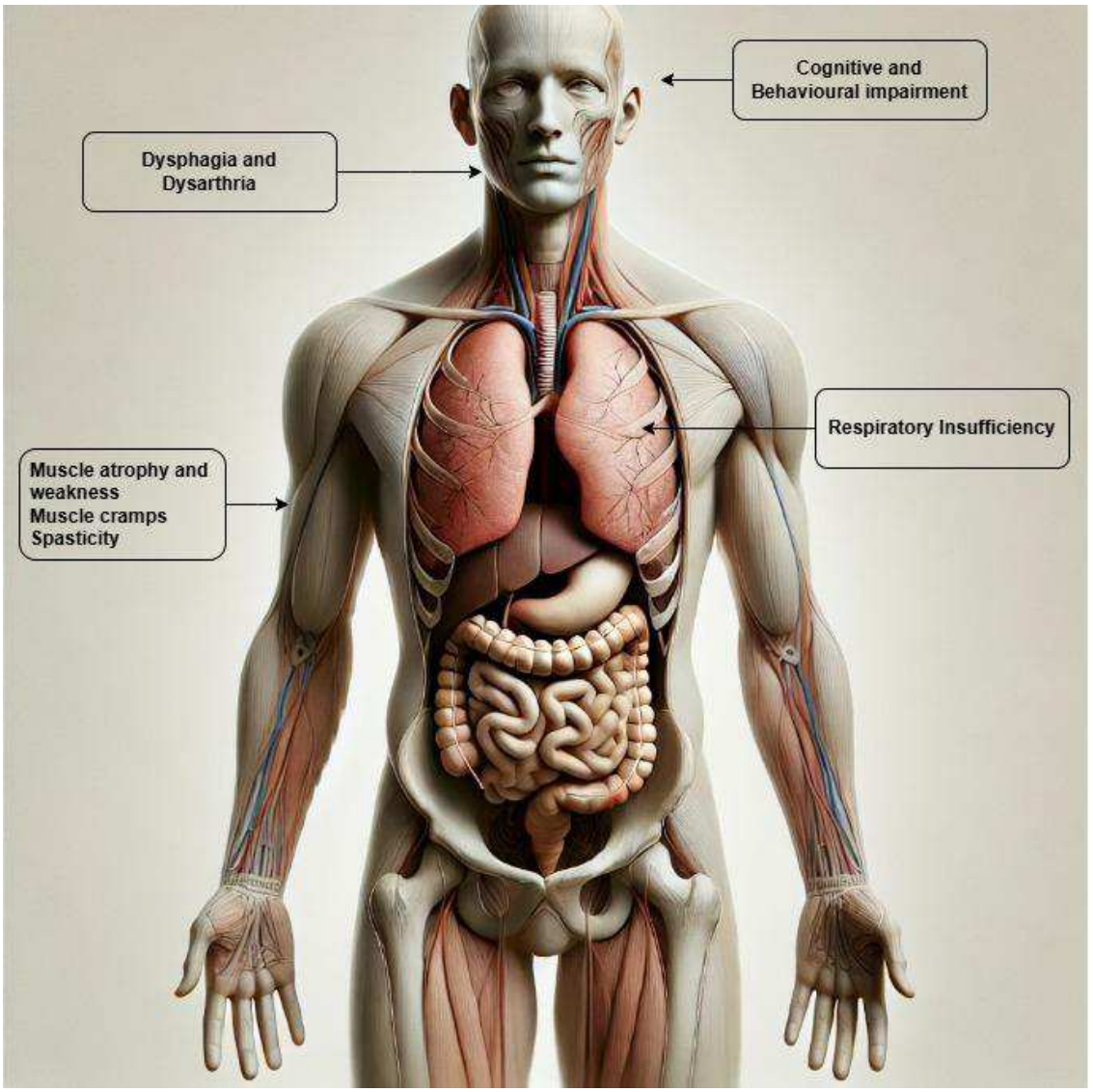

Figure 1a. Clinical manifestation of ALS.

ALS occurs in familial forms (fALS) and sporadic forms (sALS). The prevalence of sporadic forms varies between 90% and 95% of the cases. To the current knowledge, familial cases account for 5% or 10% of the cases [4]. The incidence of ALS is 1 to 2.6 over 100,000 individuals per year and 4 to 8 for the most at-risk group, which are those between 45 and 75 years old [5,6]. Men have a higher risk of developing a sporadic form of ALS compared to women [7]. However, this event is not

reflected in the familial forms. The mean age at onset is approximately between 47 and 53 years old in familial cases and 58 and 63 in sporadic cases. Juvenile (<30) and senile forms (>75) were additionally observed and reported [8]. Survival in ALS patients typically ranges from 2 to 5 years, but it depends on the patient's age and the initial site of disease onset. However, about 20% of patients survive more than 5 years, 10% at least 10 years, and 5% at least 20 years [9].

ALS could emerge in 3 different clinical forms, spinal, bulbar, and respiratory [10]. Bulbar forms account for 25% of the cases, are often characterized by initial speech and feeding difficulties, and are associated with the worst prognosis [11]. Spinal cases, on the other end, account for 75% and onset with weakness and atrophy of the arms or limbs. Respiratory ALS is only 3% and dyspnea could be the first noticeable sign, occurring before any limb weakness.

ALS risk factors and molecular bases

ALS has been associated with several environmental and lifestyle risk factors, such as air pollution, heavy metals pollution, smoking, repeated head trauma, and intense physical activity [12–14]. Despite this, the literature remains conflicting, and no clear associations with particular phenotypes of the disease have been found. At the state of the art, aging, being male, or family history are the only well-established ALS risk factors.

The molecular bases of ALS are an interesting and challenging area of study due to the complexity and heterogeneity of the disease. We currently know that the majority of ALS cases share TDP-43 proteinopathy, characterized by the abnormal aggregation of TDP-43 protein in neurons and glial cells [15]. TDP-43 is a nuclear and ubiquitary protein, transcribed from the *TARDBP* gene mRNA, regulating miRNA biogenesis and splicing, mRNA transcription and traduction, and some other stress responses. Its mislocalization and subsequent aggregation could disrupt normal cellular functions, leading to neurodegeneration. TDP-43's central role in ALS pathology makes it an intriguing target for pharmaceutical approaches. The key idea is to develop therapies that can prevent or reduce TDP-43 aggregation and potentially slow the disease progression. Despite this, TDP-43-targeting therapies failed to restore normal cellular homeostasis in humans and few drugs passed to clinical trials [16].

The Genetic Landscape of ALS

Over the past 30 years, research has demonstrated that genetics represent the primary ALS susceptibility risk. Based on our current knowledge, approximately 76% of familial and 25% of sporadic ALS cases could be explained by 32 genes ALS-associated (Figure 1b). Whereas, pathogenic mutations in the 4 main ALS-associated genes account for 60% of familial and 11% of sporadic cases [17]. *SOD1* was the first gene associated with ALS in 1991 by a parametric linkage analysis [18]. *SOD1* encodes for superoxide dismutase 1, an enzyme that converts superoxide (O_2^-) into less dangerous hydrogen peroxide (H_2O_2) and oxygen (O_2). The toxic gain of function of the protein causes neuronal cell death by excitotoxicity, non-cell-autonomous toxicity of neuroglia, oxidative stress, mitochondrial dysfunction, and axonal transport disruption [19]. Additionally, misfolded protein seems to play a role in some sporadic forms. Missense mutations are prevalent with autosomal dominant inheritance, except for p.Asp91Ala with recessive inheritance in the Scandinavian population [20]. Approximately 25% of fALS and 2% of sALS cases could be explained by 200 *SOD1* variants [21]. Recently, therapy based on antisense oligonucleotide (ASO) was developed for *SOD1* mutation carriers. The formulated drug, Tofersen [22], reduces the aberrant protein, consequently slowing the progression of the disease. Clinical studies demonstrate a significantly lower level of neurofilament light chain in the cerebral spinal fluid (CSF) in patients treated with Tofersen compared to controls [23].

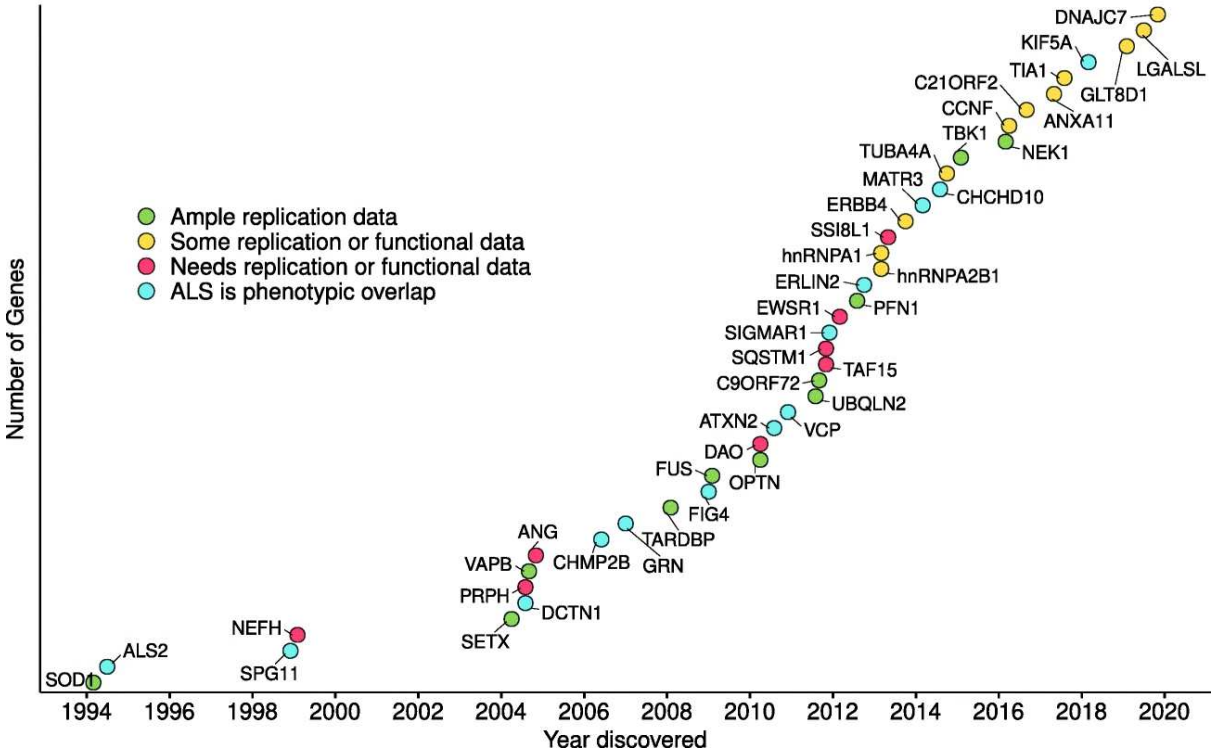

Figure 1b. Genes ALS associated discovered since 1994. Image modified from Jenna et al. Genetics of Amyotrophic Lateral Sclerosis. Curr Genet Med Rep 8, 121–131 (2020). https://doi.org/10.1007/s40142-020-00194-8

The previously cited *TARDBP* gene was discovered in 2008 in relation to the cytoplasmic aggregation of its protein TDP-43 in motor neurons [24]. Several studies have revealed the role of pathogenic variants in the C-terminal domain of TDP-43, also known as the glycine-rich region, in protein aggregation. In contrast, mutations in the N-terminal of the protein are not linked to any aggregation mechanism. Additionally, TDP-43 regulates splicing mechanisms, influencing the expression of various genes implicated in ALS processes [25]. Intriguingly, TDP-43 aggregation is in common with other neuropathologies such as frontotemporal dementia (FTD) and limbic-predominant age-related TDP-43 encephalopathy (LATE), demonstrating an underlying genetic continuum between similar age-related neurological diseases [26].

In 2009, parametric linkage analysis revealed the *FUS* gene as an ALS-associated gene [27,28]. Approximately 4% of the familial and 1% of the sporadic forms are explained by mutations in *FUS* [29]. The protein encoded by *FUS* in wild-type conditions regulates the signals in response to DNA damage and accomplishes transcription and stabilization of mRNA [30]. Mutations in *FUS* promote protein aggregation in neural cells similar to TDP-43. Specifically, *FUS* has a prion-like

domain that promotes pathological aggregation and an RRM domain that allows it to bind with RNA and DNA molecules. Transmission is autosomal dominant and patients carrying mutations have a premature age of onset compared to other mutants.

An important hallmark of ALS was discovered in 2011 when the hexanucleotide repeat expansion (GGGGCC) in *C9orf72* was first associated with ALS disease [31,32]. Previous studies have tried to establish the genetic cause behind the signals identified by parametric linkage analysis in locus 9p21.3-p13.3 [33]. Despite the notable efforts, all these attempts failed since the attention was initially focused on coding regions. Only afterwards southern blot analyses revealed the hexanucleotide repeat expansion nested in *C9orf72* first intron. This expansion is the first known pathogenic cause of ALS and FTD and accounts for 40% of familial and sporadic ALS cases and 25% of FTD cases. The latter percentage increases to 50% in the occurrence of ALS+FTD patients [17]. The pathogenic mechanisms of *C9orf72* have not yet been completely uncovered. Currently, 3 different models have been proposed to explain this gap in our understanding. First, *C9orf72* was found to be involved in regulating autophagy and vesicular trafficking in both neurons and glial cells. The expansion could actively decrease the expression of the protein, evoking a haploinsufficiency mechanism [34]. Second, the mRNA GC-rich sequence in *C9orf72* could fold in stable secondary structures that may originate aggregated RNA foci. These structures could contribute to cellular instability by seizing essential nuclear factors leading to neurodegeneration [35]. Third, a non-canonical mechanism of translation initiation called repeat-associated non-AUG (RAN) translation, has been found to induce dipeptide repeats from multiple reading frames in C9orf72-positive cases. These dipeptides, particularly those containing arginine, disrupt cytoskeletal dynamics and axonal transport by interacting with kinesin motor proteins, like KIF5A, and the ubiquitin-proteasome system [36]. The wildtype allele spans between 2 and 23 repeat units, individuals with >30 repeats are considered expanded and likely pathogenic carriers (notably, some pathogenic expansion can extend up to 1600 repeat units). Over the years, several studies have tried to correlate the repeat units with clinical phenotypes, such as age at onset, disease progression, and bulbar vs spinal vs respiratory. However, no significant associations or conflicting results were reported [37–39]. Additionally, some studies tried to understand if normal expansion

could act as a disease modifier in ALS-C9orf72 negative patients [40]. However, no clear associations were detected.

With the advent of next-generation sequencing (NGS), such as whole-exome sequencing (WES) and whole-genome sequencing (WGS), a considerable number of new genes were found to be associated with ALS. For instance, we could cite *NEK1*, *ERLIN1*, *KIF5A*, *HTT*, and *SPTLC1*.

NEK1 was discovered in 2016 by Kenna et al., by conducting a gene burden analysis on fALS and controls [41]. *NEK1* mutations contribute to neurodegeneration in ALS through i) disruptions in the function of microtubules of axon and cytoskeleton; ii) impairment of nuclear transport/import mechanisms. Missense and loss of function (LoF) variants in *NEK1* contribute to the onset of 2% of fALS and 2% of sALS.

ERLIN1 was identified by parametric linkage analysis on consanguineous Turkish families [42]. Interestingly, mutation carriers have shown an earlier age at onset and a slow disease progression. The gene encodes for a complex responsible for the degradation of inositol 1,4,5-trisphosphate intracellular receptor (IP3R) ion channels that lead to synaptic loss [43].

In 2018 *KIF5A* was discovered through common and rare genome-wide analysis [44]. The pathogenic mutations in this gene are responsible for less than 1% of fALS and sALS. The inheritance is autosomal dominant and the toxic gain of function is due to mutations skipping exon 27. The novel protein of 39 amino acids leads to hyperactive axonal transport [45].

HTT pathogenic CAG repeat expansions were commonly associated with Huntington's disease [46]. However, a recent study has shown an association with ALS and FTD using a large whole-genome dataset [47]. Interestingly, none of the ALS/FTD patients have shown the typical Huntington's symptoms, such as displayed chorea, and abnormal involuntary movements. Nevertheless, inclusions of TDP-43 aggregations were retrieved in post-mortem tissue. This finding reinforces the idea of a continuous spectrum underlying several different neurodegenerative diseases.

SPTLC1 represents a peculiar case, as mutations in coding sequence lead to cases of juvenile ALS, with an onset <25 years old [48]. The molecular bases are attributable to the disruption of sphingolipid metabolism in motor neuron disease. Even in this case, we are dealing with a gene that has been previously associated with other neurodegenerative diseases, such as autosomal dominant hereditary sensory autonomic neuropathy, type 1A (HSAN1A).

The role of genetic epidemiology in the study of complex traits

In the previous paragraph, we cited several statistical and epidemiological methods to analyze the genetics of complex and rare diseases. Herein, we provided a comprehensive and detailed overview of these methods. These strategies include genome-wide association studies (GWAS), epigenome-wide association studies (EWAS), and NGS approaches (WES, WGS). Additionally, various statistical methods can subsequently be applied to data generated for association analyses, including parametric linkage analysis in family studies, non-parametric linkage analysis in unrelated samples, such as Loss of Heterozygosity (LOH) analysis or Identity By Descend (IBD) on unrelated cohort, rare Copy Number Variation (CNVs) analysis, and rare single nucleotide variant or insertions/deletions analysis (RVs).

The common statistical method applied in GWAS is logistic regression for binary traits like disease status or linear regression for quantitative traits [49]. Using a linear model, for every SNP, the genotype-SNP relationship with the trait could be modeled with the equation for binary traits:

$$logit(P(trait=1))=\beta0+\beta1*snp+\beta2*covariates$$

Here, P(trait=1) is the probability of having a trait (e.g. disease status); β0 is the intercept; and β1 is the coefficient for the SNP, describing its effect on the odds of the trait. Covariates like age and sex are included to account for any confounding variables. The calculation for the odds ratio of the SNP is given as eβ1 and denotes the change in odds of a trait given each additional copy of the minor allele.

For quantitative phenotypes, the model may look as follows:

$$Y=\beta0+\beta1*snp+\beta2*covariates+\epsilon$$

Whereby Y is for the quantitative trait, β0 the intercept, β1 for the SNP coefficient, for instance, the effect on the trait, and ϵ represents the error term. β1 estimates the SNP effect size to mean change in the trait for each additional copy of the minor allele. In both cases, to correct the massive number of comparisons while testing millions of SNPs, the stringent significance threshold is set at 5×10−8. These methods have been proven to be effective for several neurological diseases, including ALS [50–52]. EWAS follows a similar approach, using methylation beta values

as a dependent variable in the linear regression model. The beta value represents the ratio between the signal intensity of the methylated probe and the sum of the intensities of the methylated and unmethylated signals. Its values could vary between 0 and 1 and are calculated using the subsequent formula:

$$\beta = M/(M + U)$$

Where M represents the intensity of the methylated probe and U of the unmethylated.

Parametric linkage analyses are family-based studies that aim to identify genetic regions associated with a trait or disease. Compared with GWAS, which are better designed to detect common risk variants, linkage studies are better at identifying genes containing rare high-penetrance risk variants. The assumptions required by these strategies include the specification of the disease model, autosomal dominant or recessive, and the penetrance of the disease within the family. The key statistic for parametric linkage analysis is the logarithm of the odds (LOD) score [53]. LOD score divides the probability of the data given that linkage exists at a particular value of the recombination fraction θ to the hypothesis of no linkage (θ = 0.5), and is defined by the equation:

$$LOD(\theta) = log10(L(\theta) / L(0.5))$$

or

$$LOD = log10(P(data \mid linkage) / P(data \mid no\ linkage))$$

where L(θ) is the likelihood of data given a recombination fraction θ, and L(0.5) is the likelihood under no linkage. LOD score above a value of 3 indicates an association between the marker and the observed trait. This method resulted in extremely robust new findings in genetics, as previously described [18,27,54].

Non-parametric linkage (NPL) approaches offer significant robustness in analyzing complex traits, particularly when the genetic model is uncertain. However, they commonly require a larger sample size compared to parametric methods [55]. Underlying the NPL hypothesis is the assumption that afflicted individuals possess alleles in linkage disequilibrium (LD) with pathogenic mutations or identical by descent (IBD) susceptibility alleles [56]. The common statistic for NPL is defined by the

S or S_{all} which measures the total amount of alleles shared by IBD and is determined by the equation:

$$S = \sum_{i<j} IBDij$$

Where i e j represents the compared individuals, and IBDij is the total amount of alleles shared by the individuals.

In this context, LOH analysis, also known as autozygosity mapping, and IBD analysis results are appropriate for identifying genetic regions that have lost variability, which may be associated with diseases [57]. Even though these techniques were widely applied, their effectiveness often suffered since disease locus mapping to huge genomic intervals (e.g., > 50 Mb) made it challenging to identify causal susceptibility variants [56,58]. However, these techniques have proven to be frequently successful in recognizing associated genes in recessive states that escape the classical screening for several disorders, including ALS [41,57,59–61].

In the last decade, NGS-based methods have greatly boosted genetic association analysis. The lower cost and capability to generate these data on a large scale have allowed the development of various analysis approaches. Herein, we have dealt only with CNVs and RVs analyses, nevertheless, they represent only a small fraction of the bioinformatics tests available, such as burden testing of rare variants or structural variants (SVs) analysis [62,63]. The analysis of rare variants allows for accurate screening of scattered signals across the genome, best achieved by filtering based on frequency in reference databases like gnomAD and pathogenicity scores from prediction tools. However, this process doesn't significantly reduce their number to identify those linked to the phenotype easily, and intronic region analysis remains complex and unexplored, making rare variant analysis more sensible in association studies like burden tests or diagnostic phases on disease-associated genes. CNVs analysis of NGS data relies on read depth when calling a duplication or a deletion. Tools based on this method assume that any deviation from a normal distribution of read lengths could harbor a CNV. Any increments in the number of reads could reveal a duplicated segment, while, decrements a deletion. ExomeDepth, a tool developed for calling CNVs specifically on WES data, computes Bayes Factors, a likelihood ratio of CNV probability to normal copy number probability. A Bayes factor near 3 could indicate moderate evidence of a CNV and values upper than 9 strong

evidence [64,65]. However, the higher rate of false positives significantly influences the interpretation of CNVs in the context of complex and rare diseases.

Research Hypothesis and Experimental Plan

The development and refinement of statistical and bioinformatic pipelines have provided the flourishing of a notable number of outcomes related to genetics and association analysis. Despite these significant advances, currently, there is a lack of comprehensive methods that aggregate the various results generated by these different approaches. This limitation could lead to a narrow and limited view of the findings, as each technique often presents data from a unique perspective, thereby missing the broader context and the interconnectedness of the information that would offer a clearer global picture. Currently, the main approaches for this purpose focus on performing meta-analyses on GWAS, integrating information from multi-omic data, or using colocalization methods [51,66]. Nevertheless, these methods do not consider the integration of all possible analyses that could be performed on the data itself. To overcome these pitfalls, we developed **Gen**etic m**Ul**tI a**n**alysis aggr**E**gator (GenUInE), a Python tool that aims to converge multiple signals across the genome, deriving from different analyses. GenUInE takes advantage of the standard BED files as input, characterized by 3 columns indicating the genomic locations, and generates a final enriched matrix with a binary representation of the presence or absence of the input interval within several predefined genetic windows. Finally, GenUInE computes the probability of observing signals within each genetic window based on input data. To demonstrate the validity of our tool, we tested GenUInE functions using data from the sporadic ALS cohort provided by the Laboratory of Neuroscience of IRCCS Istituto Auxologico Italiano. The adopted workflow first required the generation of results from multiple analyses. In particular, we previously performed the following steps:

1. LOH analysis on genotyping data to extract ROHs.
2. IBD segment extractions using genotyping data.
3. Calculation of epivariation using EWAS data.
4. Extraction of SNVs from WGS and WES data.
5. Extraction of INDELs from WGS and WES data.
6. CNVs calculation on WGS and WES data.

After these steps, we converted the results from each analysis into BED files required from GenUInE and analyzed the final matrix. With GenUInE we are proposing a novel and multiperspective method to explore and prioritize genetic data generated from different analyses. We believe that this strategy could dramatically increase our ability to detect hotspots in the genomes of patients suffering from rare and complex diseases, such as ALS. Moreover, GenUInE is versatile and completely open source and can be applied to various diseases or other types of analyses, independently of the nature of the analyses themselves. Results and generated code were made available on the public repository and GitHub under MIT license.

Materials and methods

Cohort and available data

Over the years the laboratory of neuroscience of IRCCS Auxologico has collected a large cohort of 4,000 well-characterized ALS patients of Italian descent. Samples were enrolled following the El Escorial revised criteria [67]. The screening of this cohort for ALS-causing mutations has provided invaluable insight into the genetic epidemiology of ALS in Italy [68,68–70] Specifically, we generated genotyping data from 5,500 ALS patients and 4,000 controls using different Illumina arrays (including Illumina 660W-Quad BeadChips and Global Screening Arrays). Furthermore, a subsample of 61 sporadic ALS cases and 61 controls underwent methylation screening by methylation array. Our laboratory could additionally count on ~200 ALS WGS data and ~350 ALS WES data already generated. Finally, our group is directly involved in the ALS Compute project, an international collaboration aimed at identifying genetic risk factors in ALS through WGS. ALS compute already collected a total of 15,000 WGS, ~12,000 ALS cases and ~3,000 controls, and is expected to increase to more than 50,000 genomes in the next few years.

Main data types and formats

Several data types are required in bioinformatics workflows. The main formats used in our analyses were PEDMAP, BED, and VCF. PedMap is the most used format for storing genotyping data and consists of two separate files: the PED file and the MAP file. The PED file is a tab-separated file that contains 6 mandatory columns that represent in order, Family ID, Individual ID, paternal ID, maternal ID, Sex, and phenotype. Genotypes are stored from the seventh column onward. The MAP file includes information about the markers, such as the genomic location and unique identification number. BED files are tab-delimited text files that outline the genetic location of specific regions. Each line has 3 mandatory columns that contain i) chromosomes, ii) the starting base position, and iii) the ending base position. VCF (Variant Call Format) file is a tab-separated text file used in bioinformatics to store and share genetic variant information, containing a header with metadata and definitions for data columns, and a data section that describes the chromosomal

positions, reference and alternate alleles, quality scores, filter statuses, additional annotations, and genotype data for multiple samples.

Genotyping arrays and quality control

Genomic DNA was extracted from peripheral blood using the Wizard Genomic DNA Purification Kit (Promega, Madison, WI, USA). SNP genotyping was conducted with Human 660W-Quad BeadChips and Global Screening Arrays on the HiScan platform (Illumina, San Diego, CA, USA) following the manufacturer's instructions. The resulting SNP data were analyzed using GenomeStudio software (Illumina) and exported in PEDMAP format. Multiple rounds of quality control were performed using the Plink software [71]. Data were cleaned according to the standard procedure:

1. **Sex genotype-phenotype mismatches.** All samples in which the genotypic sex does not correspond to the phenotypic were removed (all samples tagged with “PROBLEM”).
2. **SNP missing and call rate.** All SNP with a genotyping rate lower than 99% and at least collected in 95% of samples were excluded from the analysis.
3. **Relatedness analysis**. Discarded all samples duplicated or related (P_HAT > 0.5).
4. **Population analysis.** Samples were compared to a reference panel (HapMap2) of SNP divided into different batches corresponding to specific populations (e.g. European, African…etc.) using Eigenstrat software. Individuals who could not be ascribed to the European block were excluded from the analysis.

The workflow and pipeline are described in Figure 2a.

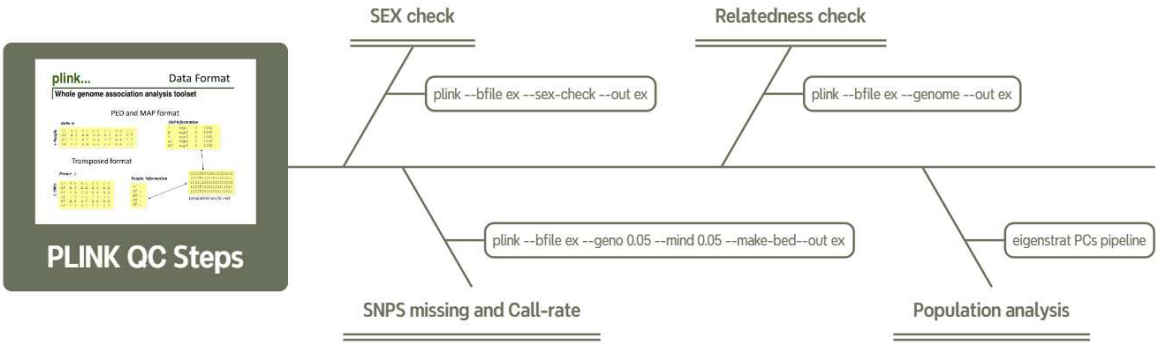

Figure 2a. Graphical pipeline summarizing the QC steps performed for genotyping data

Homozygosity mapping

Plink could rely on several functions for multiple analytical scenarios, and specifically, is considered robust for studying LOH and ROHs. The PLINK algorithm for detecting ROHs in genotyping data employs a scanning window approach and is evoked by the *--homozyg* function. This scanning window is defined by the number of SNPs determined by the user (*--homozyg-window-snp*), with a maximum number of heterozygous SNPs and of missing SNPs. Each individual's genome is controlled subsequently, scoring each SNP based on the proportion it appears in a homozygous window. After this step, genome-wide segments of homozygous SNPs are identified using a specific threshold (*--homozyg-window-threshold*). For instance, if a window size contains 100 SNPs, using a threshold of 0.05 each SNP must appear in at least five homozygous windows to be considered part of a segment. The last step implies additional conditions for these homozygous segments to identify the final ROH segments. The maximum interval between two SNPs in a segment and the maximum number of heterozygous SNPs allowed in the final ROH segment are evaluated by setting *--homozyg-gap* and *--homozyg-het* parameters. Then ROH segments exceeding the selected thresholds are split and re-evaluated, leading to ROH segments smaller than the scanning window size. Finally, the SNP density, ROHs minimum length, and number of SNPs per segment are checked respectively with *--homozyg-density, --homozyg-kb,* and *--homozyg-snp*. Only ROHs that meet all these criteria were considered in the downstream analysis. In our analysis, we selected ROHs using default not-stringent parameters, specifically:

--homozyg-snp: ***100 SNPs***
--homozyg-kb: ***>=1,000 kb***
--homozyg-density: ***50 kb/SNP***
--homozyg-gap: ***1,000 kb***
--homozyg-window-snp: ***50 SNPs***
--homozyg-window-missing: ***5 missing calls maximum***
--homozyg-window-threshold: ***0.05***

The resulting ROHs were extracted and converted into BED format. LOH analysis was performed on the inbred population. The inbred population represents the offspring from consanguineous marriages and was estimated using Plink on genotyping data. Wright F coefficient was calculated considering the expected and the observed number of homozygous SNPs, all samples with a threshold over 0.05 were retained in the analysis.

Identical By Descend (IBD) regions

IBD regions were inferred using the KING software [72]. KING accepts the PedMap format as input and can calculate the IBD regions between pairs of individuals up to a certain established degree. The parameter *--ibdseg* was set as mandatory to call the function on the input genotypes matrix. The resulting tab-separated file contains the genomic location of the inferred segment or in alternative the starting and ending SNPs. Additionally, KING with the parameter *--rplot* could produce a summary plot of the inferred regions as described in Figure 2b. The generated data was subsequently converted into BED format.

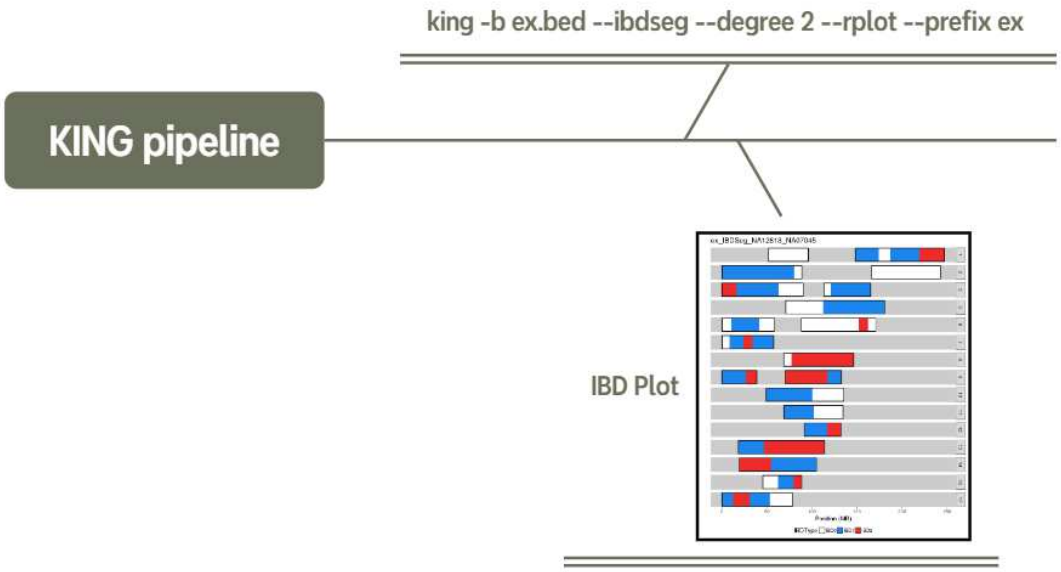

Figure 2b. KING pipeline for IBD segments estimation and visualization.

Methylation analysis

Genomic DNA (gDNA) was extracted from peripheral blood using the Wizard Genomic DNA Purification Kit (Promega). Quality control (QC) and quantification were confirmed by visualizing gDNA on a 1% agarose gel electrophoresis and using a NanoPhotometer Pearl (Implen GmbH). Bisulfite conversion was performed with the EZ DNA Methylation Kit (Zymo Research Corporation). The samples were then analyzed using the Illumina HumanMethylation450 array, following the manufacturer's best practices. Quality control (QC) of probes was initially estimated using the ChAMP package [73], with the following criteria for filtering: (1) probes with a detection p-value above 0.01, (2) probes with a beadcount <3 in at least 5% of samples, (3) probes not located in CpG sites, (4) probes flagged by Zhou et al. [74], and (5) probes situated on X and Y chromosomes. Additionally, signal intensities were normalized using the SWAN normalization method from the minfi package [75]. Batch effects due to experimental variability were evaluated and adjusted using the ComBat R methods [76], with the batch group (i.e., different experiment groups) as a covariate. After this step, we calculated Stochastic epigenetic mutations (SEMs). SEM represents those CpG sites with a methylation level exceeding three times the

IQR below the 25th percentile (Q1–3 × IQR) or three times the IQR above the 75th percentile (Q3 + 3 × IQR) [77–80]. SEM calculation was the key step for assessing the presence of epivariations. Epivariations are regions that exhibit an abnormal methylation pattern, significantly enriched in epimutations [81,82]. The calculation method, developed and validated by our laboratory, involves a sliding window on the annotated genome which uses a hypergeometric distribution to assess SEM enrichment. Then an associated p-value is generated individually for each window. Our R package for SEM calculation is available at DOI 10.5281/zenodo.3813234. The genomic coordinates of Epivariations were finally converted and exported in BED format.

WGS workflow and quality control

The genomic DNA of the ALS individuals was extracted from whole blood according to standard protocols. DNA concentration and quality were assessed using a NanoDrop spectrophotometer and agarose gel electrophoresis. Selected samples underwent WGS on the Illumina NovaSeq platform, with a mean coverage of 30x. Raw reads were processed with our custom pipeline optimized following Broad Institute's best practices. In particular, we aligned FASTQ files to the reference genome (GRCh37) into SAM(Sequence Aligned Map) files using Burrows-Wheeler Aligner [83]. SAM files were subsequently converted to their binary representations, i.e. BAM files (Binary Aligned Map). Indexes files (BAI) were generated to call back sequences to the reference genome quickly. Deduplication and recalibration steps were also performed to add reliability to the final alignment. instrumentations can make technical errors in the sequencing step, recalibration is, therefore, a crucial step that accounts for these systematic errors leading to better quality reads. We performed the variant calling step using the GATK [84] tool and the HaplotypeCaller algorithm on BAM files. HaplotypeCaller relies on a local realignment of the reads before calling the variants. This led to a low rate of false negatives in the final dataset. Recalibration and hard-filtering were additionally obtained using variant quality score recalibration (VQSR). This method, based on machine learning, employs highly validated datasets (e.g., 1000genomes, HapMap, omni) to build a subset of the input variants (true positive) as the “train” set. Good and bad variants were recognized based on their profiles on this “train” dataset. Subsequently, the

rules learned from the “train” dataset are applied to other sites for filtering purposes. Quality control was conducted at the on-site level, and only variants tagged with “PASS” were retained in the final dataset. The resulting dataset was additionally annotated using ANNOVAR with the information provided by gnomAD, 1000genomes, ClinVar, InterVar, and CADD. Extra-annotation was obtained by adding columns with:

- Frequency of variants inside the cohort.
- Frequency of variants in homozygous state within the ALS cohort.
- DP (Read Depth) list of all observed variants.

Figure 2c summarizes the entire workflow of this step.

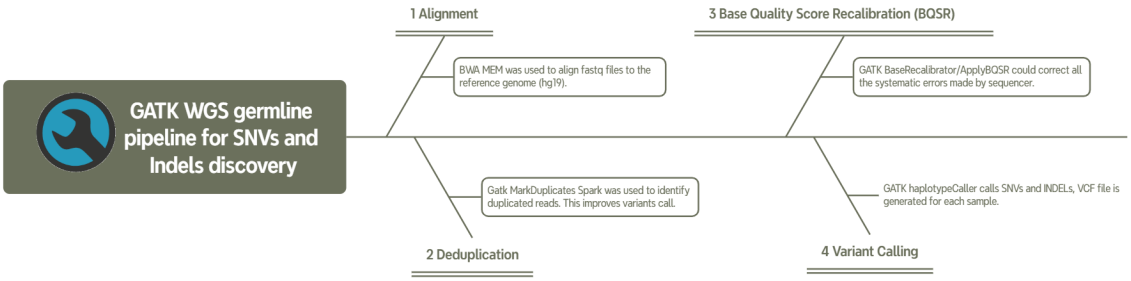

Figure 2c. GATK germline pipeline for SNVs and INDELs discovery.

Rare Variants

We extracted rare homozygous variants from the WES and WGS annotated VCFs datasets. GnomAD frequencies [85], calculated for the whole WES and WGS populations, were used to evaluate single nucleotide variants (SNVs) and insertions/deletions (INDELs). A stringent threshold of minor allele frequency (MAF<1/100,000) was set to retain only extremely rare recessive events. All the resulting variants were then converted to genomic location and stored in BED files.

Copy Number Variants (CNVs)

Rare Copy Number Variants (CNVs) were generated using the ExomeDepth R package [65]. ExomeDepth was chosen for its excellent performance in terms of specificity and sensitivity [86]. This package, developed for WES or WGS data, relies on read depth to estimate the presence or absence of CNVs. The algorithm first calculates the number of reads aligning to each reference genome position. Subsequently, the “binning” process is initialized: the genome is divided into non-overlapping bins defined by the user, and then the total read depth for each bin is calculated by summing the read depths of all positions within the bin. In our analyses, we preferred the default value of 10,000 bins. After this step, ExomeDepth depth normalizes the results based on GC and calculates the probability of having a CNV. Since ExomeDepth could calculate rare CNVs only using batches of 10 samples at a time, we decided to divide our population into randomized groups. The results of CNV calling were then converted into BED files.

Over Representation Analysis and statistical analysis

ORA was performed on ShinyGO[87] using KEGG as a functional database and a False Discovery Rate (FDR) threshold of 0.1. Summary statistical analysis was performed on Python (version 3.8).

Multi-analysis tool

We developed GenUInE (Genetic mUltI analysis aggrEgator) to converge genetic signals from multiple analyses. GenUInE is a Python tool that analyzes genomic intervals by generating a binary representation of their presence or absence within defined genetic windows on a reference genome. After the initialization, GenUIne takes the path to the input files directory and reference genome in BED format. BED files must be alphanumerically sorted, with a “chr” prefix before each chromosome. The tool includes static methods that apply rules to analyzed ranges:

1. *range_sequence*, which creates a list of genomic range tuples based on the window size
2. *range_subset*, which checks if one range is entirely contained within another.

The core method, *matrix_gen*, reads the reference genome, splits it into defined intervals (default of 10,000 bp), and then compares them with those in the BED files to build a binary matrix of presence/absence. This matrix indicates whether each genomic interval from the input BED files is present in some windows split from the reference genome. Finally, the *apply_stat* method calculates the win probability for each column, where each input column represents a single analysis or BED file, defined by the number of positive windows over all existing ones in the reference genome. Next, the binary presence/absence values are replaced with the win probability if the value is 1, and with 1 if the value is 0. Finally, the product of these independent probabilities is calculated by multiplying values in each row (i.e., in each window) and stored in a new column. The resulting values thus indicate the combined probability of having a signal in a given window of the reference genome, given the input analyses. The tool additionally provides a weighted score that emphasizes the low probabilities and accounts for the number of input analyses. The adopted formula was:

$$S = \alpha * (-10 * \log(comb\_prob)) + \beta * (\Sigma\ rows / N)$$

Where "*comb_prob*" represents the combined probability, "*Σ rows*" represents the summation of binary values for each window, and *α-β* are the assigned weights (default to 0.5).

At the end of the process, all results are exported into a tab-separated file. The workflow is summarized in Figure 2d.

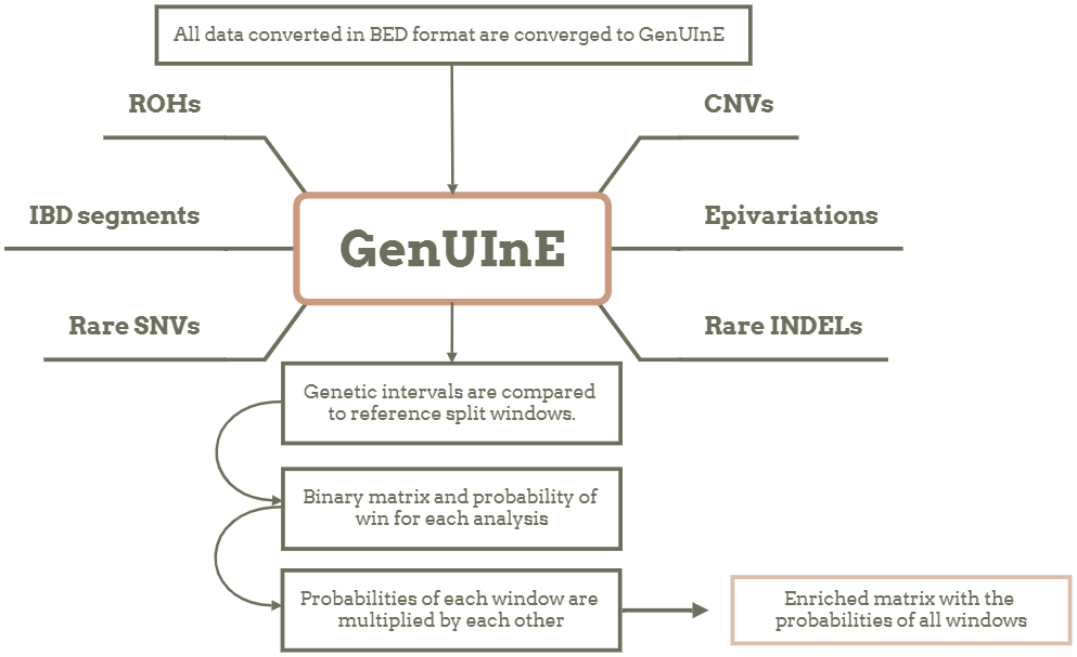

Figure 2d. How GenUInE works for calculating probabilities for each genomic window.

Here is provided an example of GenUInE on work:

```
#Input parameters:
path = "/path/to/bed/files/"
output_name = "GENOMIC_ANALYSIS"
genome = "hg19.bed"
window_size = 10000

#Create an instance of the GenUInE class
genomic_analysis = GenUInE(path=path,
 outputName=output_name, genome=genome, window=window_size)

#Generate the binary matrix
genomic_analysis.matrix_gen()

#Output: A file named 'ENRICHED_MATRIX_GENOMIC_ANALYSIS.tsv' will be created in the working directory.
```

Data and code availability

All the developed code was deposited on github (https://github.com/albertobrusati/genuine/) under the MIT license. Data will be submitted to public repositories following the best practices.

Results

Cohort summary statistics

We have collected and organized all currently available phenotypic data at our research center, IRCCS Istituto Auxologico Italiano. Our comprehensive database was developed to systematically gather a wide array of information, ranging from clinical data, such as patient age, type of disease onset, phenotypic characteristics, and sex, to additional details including diagnostic delay, disease progression, cognitive status, blood test results, biomarkers, and findings from magnetic resonance imaging (MRI). Moreover, we have integrated genetic diagnostic data wherever possible to enhance the depth and breadth of our dataset. The foundational clinical information is presented in Table 1.

ALS Cohort IRCCS Auxologico	
Total ALS cases N=1,444	Females (38%)
	Males (62%)
Age	Mean=60.21 SD=12.35
Site Of Onset	Spinal-LowerLimbs (41.9%) Spinal-UpperLimbs (33.0%) Bulbar (23.3%) Respiratory (1.7%)
ALS Phenotypes	classical (52.8%) bulbar (20.4%) respiratory (1.8%) flail-arm (3.9%) flail-leg (3.0%) PLS (5.6%) PMA (4.5%) UMNP (8.1%)

Table 1. ALS Cohort clinical phenotypes from IRCCS Auxologico.

Results from single analyses

Each genetic analysis yielded the following results and statistics. The study of ROHs across individuals demonstrated a mean number of 45.3 segments per individual, with a standard deviation of 13.0073. The 95% confidence interval for the mean number of ROH segments was calculated to be between 41.2 and 49.3. Additionally, the average length of these ROH segments was 6,811 base pairs (bp), with a standard deviation of 2,755 bp. The 95% confidence interval for the mean ROH length ranged from 5,957 bp to 7,665 bp.

In the IBD segments analysis, the mean number of segments per individual was 44, with a standard deviation of 11.78. The 95% confidence interval for the mean number of IBD segments extended from 14.7 to 73.2. The mean length of IBD segments was reported as 14 bp, with a standard deviation of 10.5 bp. The confidence interval for the mean length of IBD segments was computed to be between 12.2 and 15.9 bp.

We identified a total of 271 epivariations. The mean length of these epivariations was 549 bp, with a standard deviation of 617 bp. The 95% confidence interval for the mean length of epivariations was established as ranging from 471 to 627 bp.

Furthermore, the total number of filtered SNVs identified in the dataset was 7,451, while the total number of filtered INDELs was 1,727.

Regarding CNVs, the analysis revealed a mean number of 22.6 CNVs per individual, with a standard deviation of 14.8. The 95% confidence interval for the mean number of CNVs ranged from 19.6 to 25.5. Finally, the mean length of CNVs was determined to be 583,509 bp, with a standard deviation of 1,962,515 bp. All the summary statistics are reported in Table 2.

Genomic Feature	Metric	Mean	Standard Deviation	95% Confidence Interval
ROH Segments per Individual	Number	45.3	13.0073	[41.269-49.331]
ROH Length	Length (bp)	6,811	2,755	[5,957-7,665]
IBD Segments per Individual	Number	44	11.7	[14.7-73.2]

IBD Length	Length (bp)	14	10.5	[12.293-15.927]
Epivariations	Number	271		
Epivariation Length	Length (bp)	549	617	[471-627]
Filtered SNVs	Number	7,451		
Filtered INDELs	Number	1,727		
CNVs per Individual	Number	22.64	14.8817	[19.687- 25.593]
CNV Length	Length (bp)	583,509	1,962,515	[476,522-690,496]

Table 2. Summary results and statistics from individual analyses.

Results from GenUInE

We run GenUInE using a 10,000 bp window on bed files generated from single analyses. Of 350,000 windows, 285,768 contained at least one signal from input data. We selected the first 100 windows, corresponding to those with the lowest cumulative probability, and annotated them into the resulting genes using UCSC API. We found 218 genes mapping these regions, 60 with a strong expression in cerebral tissues as shown in Figure 3b.

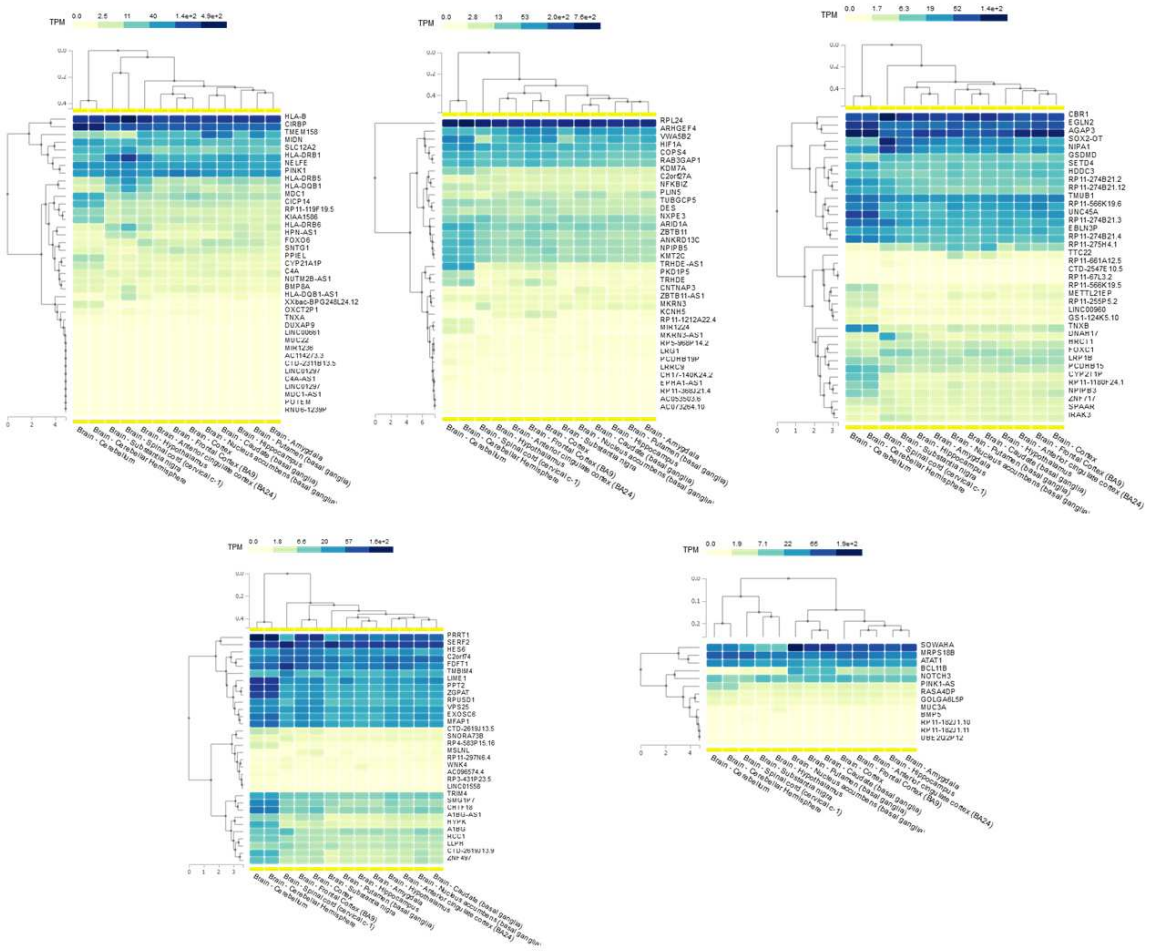

Figure 3a. Expression results in brain tissues from GTEx. Values are reported as Transcript Per Million (TPM).

We then conducted two distinct analyses on the identified genetic dataset. The first approach employed all resultant genes for an over-representation analysis (ORA) to explore the associated pathways. ORA did not reveal an enrichment in any appropriate biological pathways, with an FDR < 0.1. The second approach focused on the relationships between this genetic signature and neurodegenerative diseases. Intriguingly, the analysis showed some genes already ALS-associated (*NIPA1*), Charcot Marie Tooth (CMT) associated (*AARS1*), or associated with familial forms of Parkinson's disease (*PINK1*). Moreover, other potential candidate targets were revealed by this step, in particular, *HES6*, *SERF2*, *TRIM4*, *HIF1A*, DES, *FOXO6*, *SOX2-OT*, and *NOTCH3*. In Table 3, we reported the resulting regions and analyses containing these genes.

Genome	SNVs	EPIs	IBDs	CNVs	HOMs	INDELs	Sum	prob	score	Gene
chr2:239140000-239149999	1	1	0	0	1	1	4	3.08e-08	86.80	*HES6*
chr15:44090000-44099999	1	1	0	0	1	1	4	3.08e-08	86.80	*SERF2*
chr7:99510000-99519999	1	1	0	1	1	0	4	1.80e-06	66.46	*TRIM4*
chr14:62210000-62219999	1	0	1	1	1	1	5	5.39e-06	61.06	*HIF1A*
chr15:23100000:23109999	1	0	1	1	1	1	5	5.39e-06	61.06	*NIPA1*
chr3:181130000-181139999	1	0	1	1	0	1	4	6.65e-06	59.93	*SOX2-OT*
chr3:181450000-181459999	1	0	1	1	0	1	4	6.65e-06	59.93	*SOX2-OT*
chr1:41840000-41849999	1	0	0	1	1	1	4	1.15e-05	57.18	*FOXO6*

chr1:20960000-20969999	1	0	0	1	1	1	4	1.15e-05	57.18	*PINK1*
chr19:15310000-15319999	1	0	0	1	1	1	4	1.15e-05	57.18	*NOTCH3*
chr2:220280000-220289999	1	0	1	1	1	1	5	5.39	61.06	*DES*
chr16:70280000-70289999	1	1	1	0	1	0	4	3.51	63.12	*AARS1*

Table 3. GenUInE enriched output slice. Genes are displayed based on the connection with neurodegeneration or motor neuron diseases.

Furthermore, we analyzed our results based on a weighted prioritization score. This led to the selection of only the extreme windows (score > 80), narrowing the analysis to only 3 of the previous windows mapping 10 genes (Figure 3c), including *HES6* and *SERF2*. Table 4 contains all the resulting genes from the analyses.

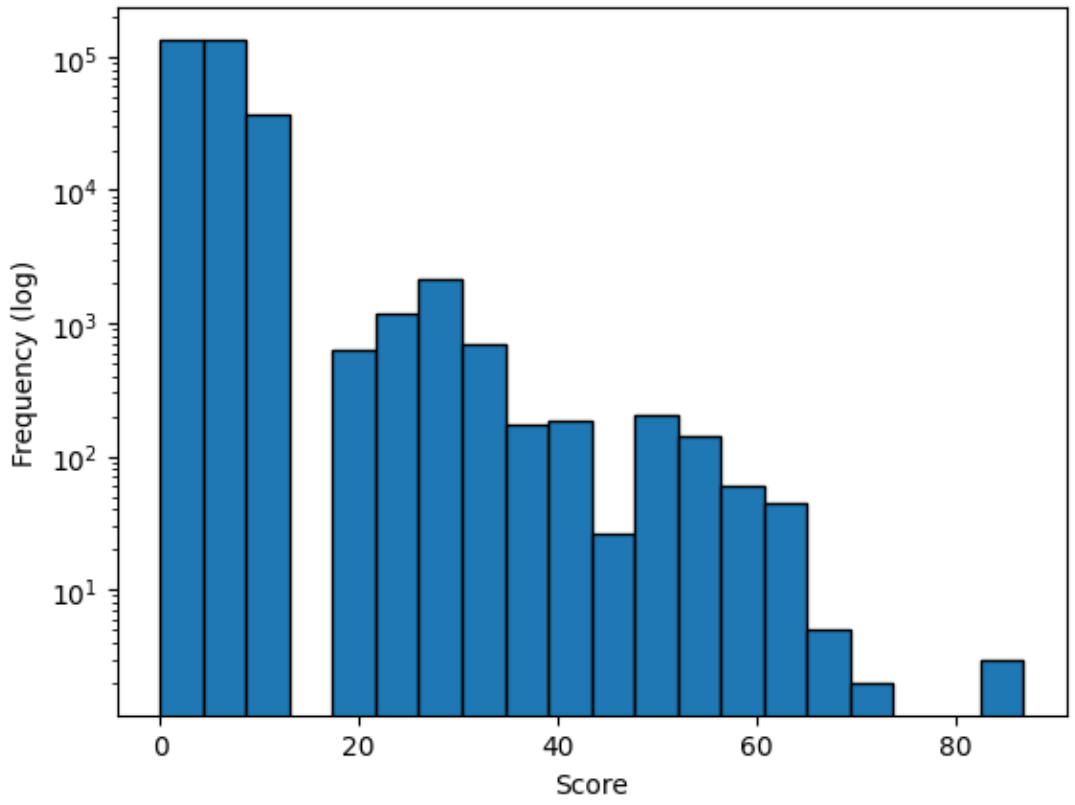

Figure 3b. Histogram representing the frequency of windows based on the calculated weight score. Y-axis represents the log transformed frequencies.

LINC02610	*ENSG00000225057*	*HES6*	*ENSG00000262560*	*SERF2*
SERINC4	*HYPK*	*MFAP1*	*FDFT1*	*ENSG00000255046*
LLPH	*ENSG00000228144*	*LLPH-DT*	*TMBIM4*	*RCC1*
SNHG3	*SNORA73B*	*LINC01558*	*ENSG00000269155*	*ENSG00000285212*
ENSG00000274769	*C2orf74-DT*	*C2orf74*	*C2orf74-AS1*	*TRIM4*
VPS25	*WNK4*	*PRRT1*	*ENSG00000285085*	*ENSG00000284954*

PPT2	*PPT2-EGFL8*	*ZGPAT*	*ENSG00000273154*	*ENSG00000274501*
LIME1	*ENSG00000273047*	*SMG1P7*	*ENSG00000291219*	*EXOSC6*
AARS1	*A1BG*	*ENSG00000268230*	*A1BG-AS1*	*ENSG00000279611*
ZNF497	*ENSG00000268049*	*MSLNL*	*RPUSD1*	*CHTF18*
PLIN5	*ENSG00000267385*	*LRG1*	*ENSG00000260978*	*MKRN3*
VWA5B2	*MIR1224*	*TUBGCP5*	*NXPE3*	*NFKBIZ*
RP11-1212A22.4	*PKD1P5*	*ENSG00000291270*	*EPHA1-AS1*	*ENSG00000229977*
COPS4	*ANAPC1P2*	*RAB3GAP1*	*LRRC9*	*PCNX4-DT*
ZBTB11	*ZBTB11-AS1*	*RPL24*	*TRHDE*	*TRHDE-AS1*
KMT2C	*ARHGEF4*	*DES*	*ENSG00000234638*	*KDM7A*
KDM7A-DT	*CNTNAP3*	*HIF1A*	*HIF1A-AS3*	*ENSG00000258964*
LINC02511	*KCNH5*	*NPIPB5*	*ENSG00000277041*	*ENSG00000260063*
ARID1A	*ANKRD13C*	*LINC03066*	*ENSG00000285578*	*C2orf27A*
ENSG00000288031	*ENSG00000280029*	*ENSG00000279983*	*PCDHB19P*	*ENSG00000290895*
PCDHB15	*LRP1B*	*FOXC1*	*EBLN3P*	*HTATSF1P2*
ENSG00000288612	*TMUB1*	*AGAP3*	*HRCT1*	*SPAAR*
RAB4B-EGLN2	*EGLN2*	*ENSG00000268797*	*CYP2T1P*	*LINC00960*
ZNF717	*MIR4273*	*ENSG00000234500*	*ENSG00000291136*	*ENSG00000277435*

ENSG00000290399	*IRAK3*	*ENSG00000290192*	*ENSG00000276548*	*NPIPB3*
NIPA1	*ENSG00000274253*	*ENSG00000259425*	*SOX2-OT*	*ENSG00000241231*
METTL21EP	*ENSG00000272542*	*DNAH17*	*HDDC3*	*UNC45A*
ENSG00000254859	*GSDMD*	*ENSG00000289161*	*TTC22*	*ENSG00000237453*
TNXB	*SETD4*	*CBR1-AS1*	*CBR1*	*ENSG00000289001*
ENSG00000242588	*ENSG00000243302*	*ENSG00000230715*	*ENSG00000229413*	*ENSG00000243679*
CICP14	*LINC02476*	*MUC22*	*TMEM158*	*AC010170.1*
HLA-B	*ENSG00000271581*	*ENSG00000293281*	*SNTG1*	*NELFE*
MIR1236	*SKIC2*	*C4A*	*C4A-AS1*	*ENSG00000290788*
CYP21A1P	*AL645922.1*	*TNXA*	*POTEM*	*ENSG00000275563*
RNU6-1239P	*SLC12A2-DT*	*SLC12A2*	*NUTM2B-AS1*	*ENSG00000280355*
ENSG00000224886	*HLA-DRB5*	*DUXAP9*	*ENSG00000286614*	*FOXO6*
FOXO6-AS1	*LINC00661*	*ENSG00000280412*	*HLA-DRB6*	*HLA-DRB1*
HLA-DQB1	*HLA-DQB1-AS1*	*MIDN*	*CIRBP*	*BMP8A*
OXCT2P1	*PPIEL*	*MDC1*	*MDC1-AS1*	*POLR1HASP*
POLR1H	*HPN-AS1*	*LINC01297*	*KIAA1586*	*PINK1*
PINK1-AS	*BMP5*	*BCL11B*	*NOTCH3*	*LINC03043*
MUC3A	*SEPTIN8*	*SOWAHA*	*ENSG00000291026*	*GOLGA6L5P*

ENSG00000291260	*ENSG00000259244*	*MRPS18B*	*ATAT1*	*UBE2Q2P12*
ENSG00000259551	*ENSG00000259302*	*RASA4DP*		

Table 4. Genes resulting from the analyses. In red, genes resulting from filtering on score value.

Benchmarks

We subjected our tool to a comprehensive benchmarking process, where the number of function calls (ncall), total execution time (tottime), and cumulative execution time (cumtime) were measured. The test was performed using all the data generated for the analyses. The total time reported includes other inherited functions and accounts for 21,831,968,982 function calls (21,831,967,164 primitive calls) over 16,202.266 seconds (4.5 hours). The performance statistics of the custom functions are summarized in Table 5. The results show that the *range_subset* function exhibited moderate performance, with over 5.4 billion calls and a total time of 1,667.106 seconds. Both *range_extreme1* and *range_extreme2* functions had very similar performance, with roughly 5.4 billion calls each, consuming just over 1,230 seconds. The *apply_stat* function was invoked only once, with a negligible time footprint, while the *matrix_gen* function, though called only once, showed the highest number of calls overall.

Function	**Number of Calls (ncalls)**	**Total Time (tottime)**	**Time per Call (percall)**	**Cumulative Time (cumtime)**	**Cumulative per Call (percall)**
range_subset	5,448,085,002	1,667.106 s	0.000 s	1,667.106 s	0.000 s

range_extreme1	5,448,065,083	1,236.243 s	0.000 s	1,236.243 s	0.000 s
range_extreme2	5,448,064,432	1,233.062 s	0.000 s	1,233.062 s	0.000 s
apply_stat	1	0.000 s	0.000 s	0.111 s	0.111 s
matrix_gen	1	216.483 s	216.483 s	16,202.240 s	16,202.240 s

Table 5. GenUInE benchmarks. Total time and number of iterations were calculated for each function.

Discussion

Genetic burden analysis still represents an intricate maze to elude, especially in the context of complex and multifactorial diseases. The best evasion attempts generally arrived from large dataset analyses, family-based approaches, or multi-omics studies. However, these approaches are not always possible due to limited resources or the low availability of high-quality data. Small sample sizes, incomplete phenotypic information, or lack of access to advanced sequencing technologies can restrain the application of large-scale or multi-omics approaches. Moreover, family-based studies may be limited by the availability of suitable pedigrees, particularly in late-onset diseases like ALS, where family history is often sporadic or poorly documented. As a result, researchers are frequently challenged to find alternative methods for highlighting genetic hotspots related to the investigated disease.

Herein, with our tool GenUInE, we presented an alternative approach to aggregate and prioritize genetic results. In particular, GenUInE consists of a cohesive framework aimed at identifying key genomic regions by integrating results from various analyses and datasets. Our method accepts BED files as input from individual analyses, such as LOH, CNV, or IBD analyses, and calculates a unified probability model across predefined genetic windows. As a result, the user could identify genomic regions related to the disease that might otherwise be overlooked by single-method analyses. Moreover, the tool estimates a score that uses a linear combination and weights the combined probability of observing a signal in a genomic window with the summation of how many analyses detected signals in that same window. In our results, we chose an equal weight of 0.5 for both the probability and summation scores, favoring a balance and avoiding possible biases. However, GenUInE allows users to adjust these weights, allowing researchers to tailor the analysis based on their specific hypotheses or research goals. In studies with higher confidence in certain types of analyses (such as WGS with high coverage), researchers could assign more weight to probability. In contrast, in exploratory analyses, summation scores might be emphasized to identify regions where multiple signals converge. The user could additionally vary the size of the examined windows. The default method uses windows of 10,000 bp, the average size of a

gene, however, it is possible to increase the size and lighten the computational load or, on the contrary, decrease the size with a more elevated computational load.

We tested GenUInE using data related to ALS as a use case, demonstrating its applicability to complex neurodegenerative diseases. The results from expression analysis on the top hits suggest effective prioritization, with a good portion of genes expressed in brain structures. On the contrary, the ORA on this subset didn't reveal any significant enrichment in relevant biological pathways. In the analysis of the genetic signature, we highlighted some known or potential candidate targets with neurodegeneration or motor neuron disorders. The expansion of the repetition unit motif GCG > 8 of *NIPA1* was associated with ALS as a risk factor [88]. Despite this, it was not replicated in all populations studied, and the association is not currently fully confirmed [89]. Interestingly, the *NIPA1* region resulted from all the input analyses except for the epivariations. This could result in contrast with our previous findings from individual analysis [14]. However, after a detailed revision, we could confirm that the region also results carrying epivariations but in a previous window ranging from 23,080,000 to 23,089,999. The *AARS1* gene has been primarily linked to peripheral neuropathies, particularly with CMT, a disorder characterized by axonal degeneration and motor dysfunction that shares some clinical features with ALS, including muscle weakness and atrophy. Some studies speculated the existence of a role in neurodegeneration caused by aminoacyl-tRNA synthetases family [90], although currently no mutations were found in ALS cases. Mutations in the *PINK1* gene are commonly associated with an autosomal form of recessive early-onset Parkinson's disease. This gene regulates the translocation of Parkin, an E3 ubiquitin ligase, to impaired mitochondria, driving their removal via mitophagy. Loss of function mutations in *PINK1* could lead to the selective loss of dopaminergic neurons in the substantia nigra [91]. The association with ALS is not direct, however, *PINK1* is known to interact with *OPTN*, a gene associated with recessive ALS forms [92]. Our analysis showed other genes pertinent to neurodegeneration or motor neuron diseases. Moreover, *HES6* and *SERF2* genes resulted in a higher score value. *HES6* and *NOTCH3* are part of the notch signaling pathway, mutations in these genes could lead to neurodegenerative responses. Interestingly, our analysis highlighted rare homozygous SNVs and INDELs in both the analyzed genes. *SERF2* is primarily involved in protein aggregation processes [93]. No explicit links with ALS are currently known, but the role in aggregation could be a promising marker. *TRIM4* can be seen

in a continuum with the previous gene, being part of the *TRIM* family, which is involved in ubiquitination and protein degradation. The disruption of these processes can subsequently lead to the accumulation of toxic protein aggregates. However, its role is mostly associated with cancer and not with neurodegeneration itself [94]. *HIF1A* is a transcription factor that responds to stress conditions such as low oxygen levels. This gene was demonstrated to be important for spinal motor neuron survival in ALS mice models after the exposition to hypoxic conditions [95]. Similar to the previous gene, *FOXO6* is involved in oxidative stress responses and its dysregulation can be hypothesized to lead to neuronal cell death [96]. *SOX2-OT* is a long non-coding RNA highly expressed in PD patients. Its role is not fully understood, but recent work has suggested its role in regulating miR-942-5p expression which in turn regulates nuclear apoptosis-inducing factor 1 (*NAIF1*) [97]. Interestingly is also the case with the *DES* gene. Desmin is a type of intermediate filament protein that primarily provides structural support to muscle cells. Autosomal recessive bi-allelic mutation in desmin has been associated with cardiomyopathies and myopathies in an Italian inbred patient [98]. This may also lead us to recognize a potential pertinence with ALS, having used rare variants in homozygosity state or ROHs as input individual analyses. Despite the promising results deriving from this genetic signature, it's crucial to emphasize that this method could highlight all those regions not considered by single analyses. In this sense, the use of multiple analytical approaches as input could lead to an increased refinement of the results. In the context of complex diseases such as ALS, this could be achieved by integrating several data from the literature.

We are not currently aware of any other method similar to what we are proposing with GenUInE. Other tools already have implemented a genomic grouping function. For instance, ShinyGo allows users to visualize regions overrepresented by the presence of multiple genes [87]. Nevertheless, ShinyGO was designed for a different purpose, particularly for gene ontology and gene-set enrichment analyses. Alternatively, it is possible to combine multiple BED files using the multiintersect function provided by BEDTools. This function allows the study of overlapping bases between different BED files. However it does not involve the division of the reference genome into windows, nor the calculation of a cumulative probability and a score to prioritize the results.

We are also conscious of the limitations that GenUInE may suffer. Dependence on the quality and coverage of input data could inflate or deflate the analysis results.

Our study applied the tool to a relatively well-characterized ALS dataset. However, its performance may be less optimal when dealing with datasets that contain incomplete or noisy data. Additionally, the sample size could affect the statistical power of this method, leading to less effective results. Our benchmarks demonstrated the satisfactory performance of the tool using real data. Nevertheless, the analysis of several individual analyses could lead to an extensive amount of time. A future perspective will be the optimization of the computational performance of the tool by reducing the number of unnecessary iterations. Moreover, our methodology adopts a fixed window size approach when analyzing the genome. While this strategy proved to be effective when analyzing large genomic regions, it may dilute the signal from contiguous windows not captured within a fixed interval. To address this shortcoming, a potential improvement for future versions of the tool will be the incorporation of a sliding window approach. This will allow better detection of signals deriving from complex genomic regions that may escape with the current strategy. Another potential pitfall may result from the fact that many analyses could not be fully independent of each other, resulting in an overestimation of certain genomic hotspots. In the future implementation, we are planning to set up different strategies that could partially mitigate this restrain by assigning a different weight to the summation value.

In conclusion, with GenUInE we are proposing a novel framework integrating multiple genetic signals into a unique and cohesive model. This method could be applied to several scenarios, in particular for rare diseases with complex genetics and partially unknown onset. Furthermore, being GenUInE an open-source project, collaboration and innovation in the scientific community could be enhanced, making it a valuable support for genetic research.

Bibliography

1. Eisen, A., Vucic, S. & Mitsumoto, H. History of ALS and the competing theories on pathogenesis: IFCN handbook chapter. *Clin. Neurophysiol. Pract.* **9**, 1–12 (2023).
2. Chiò, A., Silani, V., & Italian ALS Study Group. Amyotrophic lateral sclerosis care in Italy: a nationwide study in neurological centers. *J. Neurol. Sci.* **191**, 145–150 (2001).
3. Zayia, L. C. & Tadi, P. Neuroanatomy, Motor Neuron. in *StatPearls* (StatPearls Publishing, Treasure Island (FL), 2024).
4. Barberio, J., Lally, C., Kupelian, V., Hardiman, O. & Flanders, W. D. Estimated Familial Amyotrophic Lateral Sclerosis Proportion. *Neurol. Genet.* **9**, e200109 (2023).
5. Ryan, M., Heverin, M., McLaughlin, R. L. & Hardiman, O. Lifetime Risk and Heritability of Amyotrophic Lateral Sclerosis. *JAMA Neurol.* **76**, 1367–1374 (2019).
6. Logroscino, G. *et al.* Incidence of amyotrophic lateral sclerosis in Europe. *J. Neurol. Neurosurg. Psychiatry* **81**, 385–390 (2010).
7. Ingre, C., Roos, P. M., Piehl, F., Kamel, F. & Fang, F. Risk factors for amyotrophic lateral sclerosis. *Clin. Epidemiol.* **7**, 181–193 (2015).
8. steckinsights. Juvenile ALS: How Common is ALS in Your 20s? *Target ALS* https://www.targetals.org/2022/10/27/juvenile-als-how-common-is-als-in-your-20s/ (2022).
9. Morris, J. Amyotrophic Lateral Sclerosis (ALS) and Related Motor Neuron Diseases: An Overview. *Neurodiagnostic J.* **55**, 180–194 (2015).
10. Masrori, P. & Van Damme, P. Amyotrophic lateral sclerosis: a clinical review. *Eur. J. Neurol.* **27**, 1918–1929 (2020).
11. Moura, M. C. *et al.* Prognostic Factors in Amyotrophic Lateral Sclerosis: A Population-Based Study. *PLoS ONE* **10**, e0141500 (2015).
12. Seelen, M. *et al.* Long-Term Air Pollution Exposure and Amyotrophic Lateral Sclerosis in Netherlands: A Population-based Case–control Study. *Environ. Health Perspect.* **125**,

097023.

13. Swash, M. & Eisen, A. Hypothesis: amyotrophic lateral sclerosis and environmental pollutants. *Muscle Nerve* **62**, 187–191 (2020).
14. Brusati, A. *et al.* Exploring epigenetic drift and rare epivariations in amyotrophic lateral sclerosis by epigenome-wide association study. *Front. Aging Neurosci.* **15**, 1272135 (2023).
15. Suk, T. R. & Rousseaux, M. W. C. The role of TDP-43 mislocalization in amyotrophic lateral sclerosis. *Mol. Neurodegener.* **15**, 45 (2020).
16. Hayes, L. R. & Kalab, P. Emerging Therapies and Novel Targets for TDP-43 Proteinopathy in ALS/FTD. *Neurotherapeutics* **19**, 1061–1084 (2022).
17. Akçimen, F. *et al.* Amyotrophic lateral sclerosis: translating genetic discoveries into therapies. *Nat. Rev. Genet.* **24**, 642–658 (2023).
18. Siddique, T. *et al.* Linkage of a gene causing familial amyotrophic lateral sclerosis to chromosome 21 and evidence of genetic-locus heterogeneity. *N. Engl. J. Med.* **324**, 1381–1384 (1991).
19. Hayashi, Y., Homma, K. & Ichijo, H. SOD1 in neurotoxicity and its controversial roles in SOD1 mutation-negative ALS. *Adv. Biol. Regul.* **60**, 95–104 (2016).
20. Andersen, P. M. *et al.* Autosomal recessive adult-onset amyotrophic lateral sclerosis associated with homozygosity for Asp90Ala CuZn-superoxide dismutase mutation. A clinical and genealogical study of 36 patients. *Brain J. Neurol.* **119 (Pt 4)**, 1153–1172 (1996).
21. McCann, E. P. *et al.* Evidence for polygenic and oligogenic basis of Australian sporadic amyotrophic lateral sclerosis. *J. Med. Genet.* jmedgenet-2020-106866 (2020) doi:10.1136/jmedgenet-2020-106866.
22. Blair, H. A. Tofersen: First Approval. *Drugs* **83**, 1039–1043 (2023).
23. Miller, T. M. *et al.* Trial of Antisense Oligonucleotide Tofersen for SOD1 ALS. *N. Engl. J. Med.* **387**, 1099–1110 (2022).
24. Brandmeir, N. J. *et al.* Severe subcortical TDP-43 pathology in sporadic frontotemporal

lobar degeneration with motor neuron disease. *Acta Neuropathol. (Berl.)* **115**, 123–131 (2008).

25. Ma, X. R. *et al.* TDP-43 represses cryptic exon inclusion in the FTD–ALS gene UNC13A. *Nature* **603**, 124–130 (2022).
26. Besser, L. M., Teylan, M. A. & Nelson, P. T. Limbic Predominant Age-Related TDP-43 Encephalopathy (LATE): Clinical and Neuropathological Associations. *J. Neuropathol. Exp. Neurol.* **79**, 305–313 (2020).
27. Kwiatkowski, T. J. *et al.* Mutations in the FUS/TLS Gene on Chromosome 16 Cause Familial Amyotrophic Lateral Sclerosis. *Science* **323**, 1205–1208 (2009).
28. Ticozzi, N. *et al.* Analysis of FUS gene mutation in familial amyotrophic lateral sclerosis within an Italian cohort. *Neurology* **73**, 1180–1185 (2009).
29. Renton, A. E., Chiò, A. & Traynor, B. J. State of play in amyotrophic lateral sclerosis genetics. *Nat. Neurosci.* **17**, 17–23 (2014).
30. Ratti, A. & Buratti, E. Physiological functions and pathobiology of TDP-43 and FUS/TLS proteins. *J. Neurochem.* **138 Suppl 1**, 95–111 (2016).
31. DeJesus-Hernandez, M. *et al.* Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. *Neuron* **72**, 245–256 (2011).
32. Renton, A. E. *et al.* A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. *Neuron* **72**, 257–268 (2011).
33. Vance, C. *et al.* Familial amyotrophic lateral sclerosis with frontotemporal dementia is linked to a locus on chromosome 9p13.2-21.3. *Brain J. Neurol.* **129**, 868–876 (2006).
34. Webster, C. P., Smith, E. F., Grierson, A. J. & De Vos, K. J. C9orf72 plays a central role in Rab GTPase-dependent regulation of autophagy. *Small GTPases* **9**, 399–408 (2018).
35. Gitler, A. D. & Tsuiji, H. There has been an awakening: Emerging mechanisms of C9orf72 mutations in FTD/ALS. *Brain Res.* **1647**, 19–29 (2016).
36. Schmitz, A., Pinheiro Marques, J., Oertig, I., Maharjan, N. & Saxena, S. Emerging Perspectives on Dipeptide Repeat Proteins in C9ORF72 ALS/FTD. *Front. Cell. Neurosci.*

15, (2021).

37. Suh, E. *et al.* Semi-automated quantification of C9orf72 expansion size reveals inverse correlation between hexanucleotide repeat number and disease duration in frontotemporal degeneration. *Acta Neuropathol. (Berl.)* **130**, 363–372 (2015).
38. Beck, J. *et al.* Large C9orf72 hexanucleotide repeat expansions are seen in multiple neurodegenerative syndromes and are more frequent than expected in the UK population. *Am. J. Hum. Genet.* **92**, 345–353 (2013).
39. Gijselinck, I. *et al.* The C9orf72 repeat size correlates with onset age of disease, DNA methylation and transcriptional downregulation of the promoter. *Mol. Psychiatry* **21**, 1112–1124 (2016).
40. Peverelli, S. *et al.* Analysis of normal C9orf72 repeat length as possible disease modifier in amyotrophic lateral sclerosis. *Amyotroph. Lateral Scler. Front. Degener.* **25**, 207–210 (2024).
41. Kenna, K. P. *et al.* NEK1 variants confer susceptibility to amyotrophic lateral sclerosis. *Nat. Genet.* **48**, 1037–1042 (2016).
42. Tunca, C. *et al.* ERLIN1 mutations cause teenage-onset slowly progressive ALS in a large Turkish pedigree. *Eur. J. Hum. Genet. EJHG* **26**, 745–748 (2018).
43. Egorova, P. A. & Bezprozvanny, I. B. Inositol 1,4,5-trisphosphate receptors and neurodegenerative disorders. *FEBS J.* **285**, 3547–3565 (2018).
44. Nicolas, A. *et al.* Genome-wide Analyses Identify KIF5A as a Novel ALS Gene. *Neuron* **97**, 1267–1288 (2018).
45. Baron, D. M. *et al.* ALS-associated KIF5A mutations abolish autoinhibition resulting in a toxic gain of function. *Cell Rep.* **39**, 110598 (2022).
46. Donaldson, J., Powell, S., Rickards, N., Holmans, P. & Jones, L. What is the Pathogenic CAG Expansion Length in Huntington's Disease? *J. Huntingt. Dis.* **10**, 175–202.
47. Dewan, R. *et al.* Pathogenic Huntingtin Repeat Expansions in Patients with Frontotemporal Dementia and Amyotrophic Lateral Sclerosis. *Neuron* **109**, 448-460.e4 (2021).

48. Johnson, J. O. *et al.* Association of Variants in the SPTLC1 Gene With Juvenile Amyotrophic Lateral Sclerosis. *JAMA Neurol.* **78**, 1236–1248 (2021).
49. Wang, M. H., Cordell, H. J. & Van Steen, K. Statistical methods for genome-wide association studies. *Semin. Cancer Biol.* **55**, 53–60 (2019).
50. Kim, J. J. *et al.* Multi-ancestry genome-wide association meta-analysis of Parkinson's disease. *Nat. Genet.* **56**, 27–36 (2024).
51. van Rheenen, W. *et al.* Common and rare variant association analyses in amyotrophic lateral sclerosis identify 15 risk loci with distinct genetic architectures and neuron-specific biology. *Nat. Genet.* **53**, 1636–1648 (2021).
52. Wightman, D. P. *et al.* A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease. *Nat. Genet.* **53**, 1276–1282 (2021).
53. Rice, J. P., Saccone, N. L. & Corbett, J. The lod score method. *Adv. Genet.* **42**, 99–113 (2001).
54. Shojaee, S. *et al.* Genome-wide Linkage Analysis of a Parkinsonian-Pyramidal Syndrome Pedigree by 500 K SNP Arrays. *Am. J. Hum. Genet.* **82**, 1375–1384 (2008).
55. Tan, Q. *et al.* Power of non-parametric linkage analysis in mapping genes contributing to human longevity in long-lived sib-pairs. *Genet. Epidemiol.* **26**, 245–253 (2004).
56. Zhao, L. *et al.* A Rare Variant Nonparametric Linkage Method for Nuclear and Extended Pedigrees with Application to Late-Onset Alzheimer Disease via WGS Data. *Am. J. Hum. Genet.* **105**, 822–835 (2019).
57. Howrigan, D. P., Simonson, M. A. & Keller, M. C. Detecting autozygosity through runs of homozygosity: A comparison of three autozygosity detection algorithms. *BMC Genomics* **12**, 460 (2011).
58. L, K., Mj, D., Mp, R.-D. & Es, L. Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am. J. Hum. Genet.* **58**, (1996).
59. Morrow, E. M. *et al.* Identifying autism loci and genes by tracing recent shared ancestry. *Science* **321**, 218–223 (2008).
60. Alkuraya, F. S. Autozygome decoded. *Genet. Med.* **12**, 765–771 (2010).

61. Wang, S., Haynes, C., Barany, F. & Ott, J. Genome-Wide Autozygosity Mapping in Human Populations. *Genet. Epidemiol.* **33**, 172–180 (2009).
62. Kuzniar, A. *et al.* sv-callers: a highly portable parallel workflow for structural variant detection in whole-genome sequence data. *PeerJ* **8**, e8214 (2020).
63. Wu, M. C. *et al.* Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. *Am. J. Hum. Genet.* **89**, 82 (2011).
64. Marchuk, D. S. *et al.* Increasing the diagnostic yield of exome sequencing by copy number variant analysis. *PLoS ONE* **13**, e0209185 (2018).
65. Plagnol, V. *et al.* A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinforma. Oxf. Engl.* **28**, 2747–2754 (2012).
66. Caldi Gomes, L. *et al.* Multiomic ALS signatures highlight subclusters and sex differences suggesting the MAPK pathway as therapeutic target. *Nat. Commun.* **15**, 4893 (2024).
67. Brooks, B. R., Miller, R. G., Swash, M., Munsat, T. L., & World Federation of Neurology Research Group on Motor Neuron Diseases. El Escorial revisited: revised criteria for the diagnosis of amyotrophic lateral sclerosis. *Amyotroph. Lateral Scler. Mot. Neuron Disord. Off. Publ. World Fed. Neurol. Res. Group Mot. Neuron Dis.* **1**, 293–299 (2000).
68. Manini, A. *et al.* Association of the risk factor UNC13A with survival and upper motor neuron involvement in amyotrophic lateral sclerosis. *Front. Aging Neurosci.* **15**, (2023).
69. Gellera, C. *et al.* ATAXIN2 CAG-repeat length in Italian patients with amyotrophic lateral sclerosis: Risk factor or variant phenotype? Implication for genetic testing and counseling. *Neurobiol. Aging* **33**, (2012).
70. Brusati, A. *et al.* Analysis of miRNA rare variants in amyotrophic lateral sclerosis and in silico prediction of their biological effects. *Front. Genet.* **13**, 1055313 (2022).
71. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
72. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies.

Bioinforma. Oxf. Engl. **26**, 2867–2873 (2010).

73. Tian, Y. *et al.* ChAMP: updated methylation analysis pipeline for Illumina BeadChips. *Bioinforma. Oxf. Engl.* **33**, 3982–3984 (2017).
74. Zhou, W., Laird, P. W. & Shen, H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. *Nucleic Acids Res.* **45**, e22 (2017).
75. Aryee, M. J. *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinforma. Oxf. Engl.* **30**, 1363–1369 (2014).
76. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
77. Spada, E. *et al.* Epigenome Wide Association and Stochastic Epigenetic Mutation Analysis on Cord Blood of Preterm Birth. *Int. J. Mol. Sci.* **21**, 5044 (2020).
78. Gentilini, D. *et al.* Epigenome-wide association study in hepatocellular carcinoma: Identification of stochastic epigenetic mutations through an innovative statistical approach. *Oncotarget* **8**, 41890–41902 (2017).
79. Gentilini, D. *et al.* Multifactorial analysis of the stochastic epigenetic variability in cord blood confirmed an impact of common behavioral and environmental factors but not of in vitro conception. *Clin. Epigenetics* **10**, 77 (2018).
80. Gentilini, D. *et al.* Stochastic epigenetic mutations (DNA methylation) increase exponentially in human aging and correlate with X chromosome inactivation skewing in females. *Aging* **7**, 568–578 (2015).
81. Garg, P. *et al.* A Survey of Rare Epigenetic Variation in 23,116 Human Genomes Identifies Disease-Relevant Epivariations and CGG Expansions. *Am. J. Hum. Genet.* **107**, 654–669 (2020).
82. Gentilini, D. *et al.* Epigenetics of Autism Spectrum Disorders: A Multi-level Analysis Combining Epi-signature, Age Acceleration, Epigenetic Drift and Rare Epivariations Using Public Datasets. *Curr. Neuropharmacol.* **21**, 2362–2373 (2023).

83. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
84. Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinforma. Ed. Board Andreas Baxevanis Al* **11**, 11.10.1-11.10.33 (2013).
85. Chen, S. *et al.* A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* **625**, 92–100 (2024).
86. Cacheiro, P. *et al.* Evaluating the Calling Performance of a Rare Disease NGS Panel for Single Nucleotide and Copy Number Variants. *Mol. Diagn. Ther.* **21**, 303–313 (2017).
87. Ge, S. X., Jung, D. & Yao, R. ShinyGO: a graphical gene-set enrichment tool for animals and plants. *Bioinformatics* **36**, 2628–2629 (2020).
88. Tazelaar, G. H. P. *et al.* Association of NIPA1 repeat expansions with amyotrophic lateral sclerosis in a large international cohort. *Neurobiol. Aging* **74**, 234.e9-234.e15 (2019).
89. Nel, M. *et al.* Repeats expansions in ATXN2, NOP56, NIPA1 and ATXN1 are not associated with ALS in Africans. *IBRO Neurosci. Rep.* **10**, 130–135 (2021).
90. Vinogradova, E. S., Nikonov, O. S. & Nikonova, E. Yu. Associations between Neurological Diseases and Mutations in the Human Glycyl-tRNA Synthetase. *Biochem. Biokhimiia* **86**, S12–S23 (2021).
91. Quinn, P. M. J., Moreira, P. I., Ambrósio, A. F. & Alves, C. H. PINK1/PARKIN signalling in neurodegeneration and neuroinflammation. *Acta Neuropathol. Commun.* **8**, 189 (2020).
92. Braun, M. M. & Puglielli, L. Defective PTEN-induced kinase 1/Parkin mediated mitophagy and neurodegenerative diseases. *Front. Cell. Neurosci.* **16**, (2022).
93. Pras, A. *et al.* The cellular modifier MOAG-4/SERF drives amyloid formation through charge complementation. *EMBO J.* **40**, e107568 (2021).
94. Han, D. *et al.* The E3 Ligase TRIM4 Facilitates SET Ubiquitin-Mediated Degradation to Enhance ER-α Action in Breast Cancer. *Adv. Sci. Weinh. Baden-Wurtt. Ger.* **9**, e2201701 (2022).
95. Sato, K. *et al.* Impaired response of hypoxic sensor protein HIF-1α and its downstream

proteins in the spinal motor neurons of ALS model mice. *Brain Res.* **1473**, 55–62 (2012).

96. Rodriguez-Colman, M. J., Dansen, T. B. & Burgering, B. M. T. FOXO transcription factors as mediators of stress adaptation. *Nat. Rev. Mol. Cell Biol.* **25**, 46–64 (2024).
97. Guo, Y., Liu, Y., Wang, H. & Liu, P. Long noncoding RNA SRY-box transcription factor 2 overlapping transcript participates in Parkinson's disease by regulating the microRNA-942-5p/nuclear apoptosis-inducing factor 1 axis. *Bioengineered* **12**, 8570–8582.
98. Onore, M. E. *et al.* Bi-Allelic DES Gene Variants Causing Autosomal Recessive Myofibrillar Myopathies Affecting Both Skeletal Muscles and Cardiac Function. *Int. J. Mol. Sci.* **23**, 15906 (2022).

Developed code

GenUInE tool

```
import glob
import pandas as pd

class GenUInE:
    def __init__(self, path, outputName="ANALYSIS", genome="hg19.bed",
                 window=10000):
        self.path = path
        self.outputName = outputName
        self.genome = genome
        self.window = window

    @staticmethod
    def range_sequence(start, stop, step):
        """Create a list of multiple ranges based on a window"""
        result = list(zip(range(start, stop, step), range(start + step - 1, stop,
                 step)))
        if (stop - start) % step != 0:
            last_fst_elem = result[-1][-1] if result else start
            result.append((last_fst_elem + 1, stop))
        else:
            result = result[:-1]
            last_fst_elem = result[-1][-1] if result else start
            result.append((last_fst_elem + 1, stop))
        return result

    @staticmethod
    def range_subset(range1, range2):
        """Check if range1 is a subset of range2."""
        if range1 and range2:
```

```
            return range1.start in range2 and (range1.stop - 1) in range2
        return False

    @staticmethod
    def range_extreme1(range1, range2):
        return range1.start in range2

    @staticmethod
    def range_extreme2(range1, range2):
        return range1.stop in range2

    @staticmethod
    def apply_stat(df, num):
        dfcopy = df.copy()
        for val in range(1, num + 1):
            pwin = dfcopy.iloc[:, val].value_counts().get(1, 0) / len(df)
            dfcopy.iloc[:, val] = dfcopy.iloc[:, val].replace(1, pwin)
            dfcopy.iloc[:, val] = dfcopy.iloc[:, val].replace(0, 1)
        df["comb_prob_value"] = dfcopy.iloc[:, 1:num +
          1].prod(axis=1).astype(float)
        alpha = 0.5  # weight for combined probability
        beta = 0.5   # weight for summation
        df["score"] = alpha * (-10*np.log(df["comb_prob_value"])) + beta *
          (df["Summation"] / num)
        return df

    def matrix_gen(self):
        """Generate binary matrix every defined window"""

        print("Create reference intervals...\n")

        ref_raw = pd.read_csv(self.genome, sep="\t", names=["chr", "start",
          "end"])
        ref_coord = {}
```

```
for chrom, start, end in zip(ref_raw["chr"], ref_raw["start"],
  ref_raw["end"]):
    ref_coord[chrom] = GenUInE.range_sequence(start, end, self.window)

# Convert the ranges into sets for faster lookup
for e, v in ref_coord.items():
    ref_coord[e] = [range(elem[0], elem[1]) for elem in v]

df_ref = pd.DataFrame(dict([(k, pd.Series(v)) for k, v in
  ref_coord.items()]))
for col in df_ref:
    df_ref[col] = f'{col}:' + df_ref[col].astype(str)

df_ref = pd.concat([df_ref,
  df_ref.T.stack().reset_index(name='Genome')['Genome']], axis=1)
df_ref = df_ref[["Genome"]]
df_ref = df_ref[~df_ref["Genome"].str.contains("nan")]

print("Reference uploaded!\nCreate matrix with selected .bed\n")

n_bed = 0
for file in glob.glob(f"{self.path}*.bed"):
    fileName = file.split("/")[-1].split(".")[0]
    print(fileName)
    n_bed += 1
    df_bed = pd.read_csv(file, sep="\t", names=["chr", "start", "end"])
    df_dict = df_bed.groupby('chr').apply(lambda x: list(zip(x['start'],
  x['end']))).to_dict()

    # Pre-calculate ranges for df_dict to avoid recalculation in loops
    range_cache = {}
    for e, v in df_dict.items():
        df_dict[e] = [range(elem[0], elem[1]) for elem in v]
        for i in df_dict[e]:
```

```
        if len(i) > self.window:
            range_cache[i] = GenUInE.range_sequence(i.start, i.stop,
self.window)

 data = set()

 # Iterare su tutti i cromosomi comuni tra ref_coord e df_dict
 common_chromosomes =
set(ref_coord.keys()).intersection(df_dict.keys())
 for chr in common_chromosomes:
    ref_ranges = ref_coord[chr]
    bed_ranges = df_dict[chr]

    for ref_range in ref_ranges:
        found_match = False

        for bed_range in bed_ranges:
            if len(bed_range) > self.window:
                rangelist = range_cache.get(bed_range, [])
               # print(rangelist)
                if any(GenUInE.range_subset(range(r[0], r[1]), ref_range) or
                       GenUInE.range_extreme1(range(r[0], r[1]), ref_range)
or
                       GenUInE.range_extreme2(range(r[0], r[1]), ref_range)
for r in rangelist):
                    data.add(f"{chr}:{ref_range}")
                    print(f"{chr}:{ref_range}")
                    found_match = True
                    break
            else:
                if GenUInE.range_subset(bed_range, ref_range) or \
                   GenUInE.range_extreme1(bed_range, ref_range) or \
                   GenUInE.range_extreme2(bed_range, ref_range):
                    data.add(f"{chr}:{ref_range}")
```

```
                print(f"{chr}:{ref_range}")
                found_match = True
                break

        if found_match:
            continue

    df_mat = pd.DataFrame(list(data), columns=["Genome"])
    df_mat[fileName] = 1
    df_ref = df_ref.merge(df_mat, on="Genome", how="left").fillna(0)

df_ref['Summation'] = df_ref.iloc[:, 1:].sum(axis=1)
df_ref = GenUInE.apply_stat(df_ref, n_bed)
df_ref.to_csv(f"ENRICHED_MATRIX_{self.outputName}.tsv", sep='\t',
  index=False)
print("Enriched matrix created!")
```

Scientific production

1. **TMEM106B Acts as a Modifier of Cognitive and Motor Functions in Amyotrophic Lateral Sclerosis**
   A Manini et al., International journal of molecular sciences 23 (16), 9276, 2022

2. **Parkinsonian syndromes in motor neuron disease: a clinical study**
   J Pasquini et al., Frontiers in Aging Neuroscience 14, 917706, 2022

3. **Association of the risk factor UNC13A with survival and upper motor neuron involvement in amyotrophic lateral sclerosis**
   A Manini et al., Frontiers in Aging Neuroscience 15, 10679546, 2023

4. **Analysis of miRNA rare variants in amyotrophic lateral sclerosis and in silico prediction of their biological effects**
   A Brusati et al., Frontiers in Genetics 13, 1055313, 2022

5. **ExTaxsI: an exploration tool of biodiversity molecular data**
   G Agostinetto et al., GigaScience 11, giab092, 2022
   *co-first author

6. **Exploring epigenetic drift and rare epivariations in amyotrophic lateral sclerosis by epigenome-wide association study**
   A Brusati et al., Frontiers in Aging Neuroscience 15, 1272135, 2023

7. **Quantification of serum TDP-43 and neurofilament light chain in patients with amyotrophic lateral sclerosis stratified by UNC13A genotype**
   V Casiraghi et al.,Journal of the Neurological Sciences, 123210, 2024

8. **NEK1 haploinsufficiency worsens DNA damage, but not defective ciliogenesis, in C9ORF72 patient-derived iPSC-motoneurons**
   S Santangelo et al., Human Molecular Genetics, ddae121, 2024

9. **Epigenetic patterns, accelerated biological aging, and enhanced epigenetic drift detected 6 months following COVID-19 infection: insights from a genome-wide DNA methylation study**
   L Calzari et al., Clinical Epigenetics 16 (1), 112, 2024

10. **Association of APOE genotype and cerebrospinal fluid Aβ and tau biomarkers with cognitive and motor phenotype in amyotrophic lateral sclerosis**
    A Maranzano et al., European Journal of Neurology, e16374, 2024

11. **Analysis of normal C9orf72 repeat length as possible disease modifier in amyotrophic lateral sclerosis**
    S Peverelli et al., Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration 25 (1-2), 207-210, 2024

12. **Exploration and Retrieval of Virus-Related Molecular Data Using ExTaxsI: The Monkeypox Use Case**
    A Brusati et al., Viral Metagenomics: Methods and Protocols, 145-154

Acknowledgments

My deepest gratitude goes to Professor Vincenzo Silani, whose wisdom and guidance enlightened this journey.

To Professor Davide Gentilini, who taught me how to navigate the vast and intricate bioinformatics oceans.

To Professor Nicola Ticozzi, whose quiet yet dedicated support has been a source of strength.

I am especially thankful to Professor Antonia Ratti for her tenacious belief in me.

A sincere thank you to Luciano Calzari, whose example encouraged me.

To my dear friends and colleagues, both past and present, Patrizia, Claudia, Silvia, Marta, Valeria, Serena, Sabrina, Arianna, Donatella, Chiara, Erika, Enrico, and Anna, your friendship made every step of this experience feel like being at home.

Last but not least, to Anna, my family, and all my friends, thank you for always standing by my side.