Who bears the burden?

Rare-variant burden testing in sub-gene units to identify ALS hotspots

Name	Tessa Zonneveld
Student number	4358104
Master's	Bioinformatics & Biocomplexity
Supervisor	Kevin Kenna
Daily supervisor	Paul Hop
Second examiner	Jan Veldink
Date	21-10-2022

Abstract

Amyotrophic lateral sclerosis (ALS) is a progressive neurodegenerative disease with a large genetic component. Many of the variants seen in ALS patients and controls are so rare that a potential association between the variant and ALS cannot be detected with genome-wide association studies (GWAS). Instead, rare-variant burden (RVB) tests can be used, which combine the signal of all variants in a gene into one signal. A potential limitation of this technique is that the signal of a group of damaging variants can be weakened by the presence of neutral or protective variants in the same gene. Because damaging mutations might occur in different densities across the gene, this research aims to reduce the limitation of testing neutral and damaging mutations together by using two different methods of grouping variants: a functional domain-based method and a spatial clustering method based on the distances between variants. Both methods were tested on three ALS-associated genes: *SOD1*, *FUS* and *NEK1*. The patterns of damaging mutations found in previous studies of *SOD1* and *FUS* were replicated, i.e. no hotspots were seen in *SOD1*, while robust hotspots of rare and ultra-rare variants were seen in the C-terminus of *FUS*. In *NEK1*, two clusters dependent on single intermediate-frequency variants were seen. Additionally, an enrichment of damaging rare variants was found at the N-terminus of *NEK1*. While the spatial clustering method resulted in more consistent hotspots of ALS variants than the functional-domain based methods, combining both methods strengthens the evidence for hotspots and facilitates the interpretation of the significant results. All in all, this method has the potential to find new hotspots in known ALS genes or new genes that are associated with ALS.

Keywords

Amyotrophic lateral sclerosis, burden testing, rare variants, spatial clustering, functional domains

Layman's Summary

Amyotrophic lateral sclerosis (ALS) is a disease in which neurons controlling muscles die off, causing increasing muscle weakness and eventually death. For many patients the cause of their disease is unclear, amongst other reasons because most variations in their DNA, also called variants, are very rare. This makes it hard to pinpoint singular disease-causing variants. Instead of looking at single variants, rare-variant burden (RVB) tests look at the effect on ALS of all rare variants in a whole gene simultaneously. Not all variants in a gene result in a harmful effect on that gene. In some genes, damaging mutations group together in specific areas of the gene. This study tries to find these areas within genes by testing separate groups of variants in a gene, rather than all groups combined. We aim to see if we can find so-called hotspots of ALS variants that all contribute to the disease together.

By using and extending software developed by our research group to perform these RVB tests, we zoomed in on groups of variants in three genes that are known to be associated with ALS. In two of these genes (*SOD1* and *FUS*) our methods replicated previous findings

regarding the distribution of variants in those gene. On the third gene (*NEK1*) we applied our method and found an enrichment of potentially damaging variations in the beginning of the gene. These results show that selecting groups of variants in genes is a promising way of uncovering the important parts of known ALS genes and potentially finding new ALS genes.

Introduction

Amyotrophic lateral sclerosis (ALS) is a neurodegenerative disease with a lifetime incidence of two to three out of 1000 individuals (van Rheenen et al., 2021). It causes degeneration of upper and lower motor neurons, resulting in symptoms such as spasticity and muscle wasting, respectively (Hardiman et al., 2017). In around 10% of all cases, a Mendelian inheritance pattern is seen, which classifies them as familial ALS (FALS) (Lattante et al., 2020). The other cases are classified as sporadic ALS (SALS). For about 20% of all cases, a genetic cause is known for the alteration of the functioning of motor neurons (Vasta et al., 2022). The other cases are probably caused by an interaction between genetic and environmental factors (Hardiman et al., 2017; Lattante et al., 2020; Vasta et al., 2022). ALS genes can be related to various cellular properties, for example axon structure and function, vesicle transport, and protein homeostasis (Hardiman et al., 2017). The dysregulation of these properties can result in motor neuron injury and aberrant accumulation of proteins typical to ALS. An example of this is seen as a result of prolonged mislocalisation and aggregation of truncated forms of TDP-43, encoded by the gene *TARDBP*, that is observed 97% of ALS patients (Hardiman et al., 2017; Suk & Rousseaux, 2020). This could for example, lead to reduced mitochondrial function through gain-of-function mechanisms and reduced microtubule outgrowth through loss-of-function mechanisms (Hardiman et al., 2017; Suk & Rousseaux, 2020). Additionally, genetic variants in *SOD1*, present in 20% of FALS and 5% of SALS cases, result in misfolded proteins that cause dysregulations in the ubiquitin-proteasome system and oligodendrocyte degeneration (Ferraiuolo et al., 2016; Pal et al., 2020; Urushitani et al., 2002).

Since 2002, one technique used to find genes associated with a trait or disease outcome is the use of genome-wide association studies (GWAS) (Thomas et al., 2005; van Rheenen et al., 2021). In GWAS, common genetic variants are tested for a significant association with disease phenotype (S. Lee et al., 2014; Tam et al., 2019). Advantages of GWAS are the success rate of finding new variant-trait associations for various types of genetic variants and the insights it can provide into complex traits and ethnic variation in those traits (Tam et al., 2019). In ALS research, this technique indirectly led to the association of genetic variants in *C9orf72* with 30-40% of FALS cases and 7% of SALS cases (DeJesus-Hernandez et al., 2011; Majounie et al., 2012). However, several aspects of GWAS, both generally and specific to ALS, have limited its potential. For example, GWAS has a high multiple testing burden due to the large number of variants included in a single analysis (Tam et al., 2019). Additionally, the genetic architecture of ALS suggests that rare variants are more indicative of ALS risk than common variants (van Rheenen et al., 2021; van Rheenen et al., 2016). Both of these limitations are key to the development of rare-variant association studies.

Rare-variant association studies aim to analyse a section of genetic variation that is overlooked in GWAS, namely variants with a low minor allele frequency (MAF) (S. Lee et al., 2014). In order to analyse the role of rare variants in complex diseases, such as ALS, various rare-variant burden (RVB) tests have been developed (Povysil et al., 2019). At their core, RVB tests calculate the association of a whole gene, rather than single variants, with a disease by combining the information of all variants in a gene into one burden score for the gene. This is thought to be more powerful than testing single variants, because the low frequency of each individual variant makes the rare single variants very unlikely to be significantly associated with a disease. Traditional burden tests can be divided into binary collapsing and count collapsing methods. The binary collapsing approach aggregates the information of all variants of a gene fitting the chosen selection criteria by analysing the presence of any genetic variant as a predictor for phenotype, i.e. diseased or healthy (Cirulli et al., 2020; Povysil et al., 2019). That is, if there are no variants found in the gene, the score for that gene will be 0. If there are one

or more variants in the gene, the score for that gene will be 1. In the count collapsing approach, one uses the total number of variants of the gene found in a patient as a predictor for phenotype instead (S. Lee et al., 2012; Povysil et al., 2019). Not all variants found in a dataset are of interest in an RVB test. Selection criteria that can be considered for identifying qualifying variants are the MAF and functional annotations, such as loss-of-function (LOF) or missense variants (Cirulli et al., 2020; Povysil et al., 2019).

The use of the RVB tests described above has led to the discovery of several new ALS risk genes. For example, by selecting non-sense, splice-altering, or deleterious variants with MAFs below 0.1% found in all protein-coding genes, the gene *NEK1* was identified as a risk gene for ALS (Kenna et al., 2016; van Rheenen et al., 2021). It was mutated in around 3% of the ALS cases in the study, interacts with other ALS-associated proteins and plays a role in the regulation of DNA repair (Kenna et al., 2016). Another study found *KIF5A* to be related to ALS, which is involved in the axonal transport of, for example, neuro-filaments (Nicolas et al., 2018). The presence of genetic variants in *KIF5A* is correlated with a lower age of onset and higher median survival time (Nicolas et al., 2018). Interestingly, most ALS-related genetic variants in this study were found in only one domain of the protein, which is involved in cargo binding. Similarly, mutations in *FUS* are most often found in the Arg-Gly-Gly (RGG-)repeat region and nuclear localisation sequence (NLS) at the C-terminus of the gene (Kwiatkowski et al., 2009; Shang & Huang, 2016). This suggests that the harmful effects of genetic variants in ALS-related genes can be localised to sections of the gene, rather than spread across the whole gene. In fact, a study by Cooper-Knock et al. (2019) even implies that the strength of RVB tests is insufficient for detecting associations with ALS when harmful variants are concentrated in smaller sections of the gene, rather than spread throughout the whole gene. While there are RVB tests that account for the presence of a small number of causal variants amongst other variants in the same gene, these tests will not tell you where the causal variants are located.

In order to investigate how burden tests can be used to find enrichments of causal variants associated with ALS, also called hotspots, supervised and unsupervised methods have been suggested to divide genes into smaller groups of variants. On the one hand, Gelfman et al. (2019) describe a supervised approach in which redefined functional domains are used as the unit of interest instead of whole genes (Figure 1B). An advantage of this method is that if a domain present in a specific gene is highly associated with ALS, the functional annotations facilitate the biological interpretation of this result. Additionally, domains can occur in different genes. Thus, if a domain is found to be associated with ALS, other genes that contain the

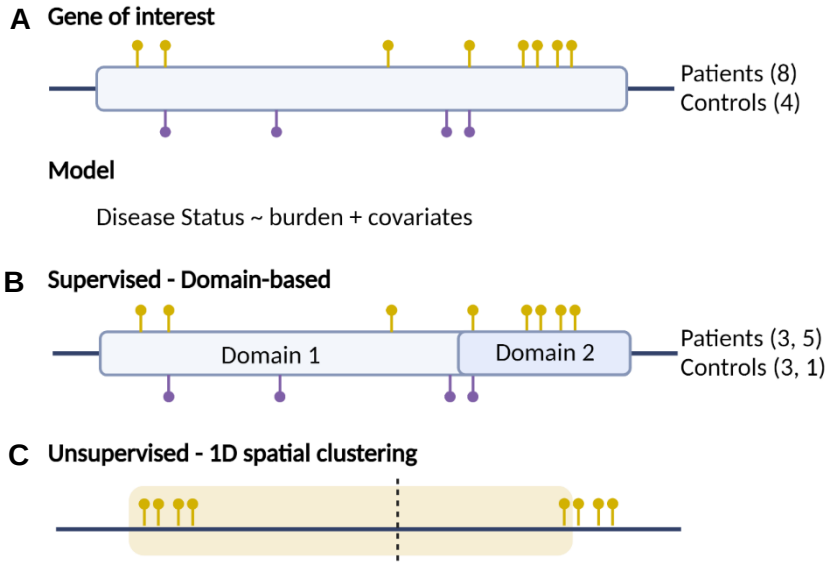


Figure 1 - Handling an unequal distribution of variants in a gene can be done in a supervised and unsupervised manner. *If a gene is expected to have hotspots of mutations related to a disease, these could be found by selecting only a part of the variants in that gene. This can, for example, be done in a supervised manner by selecting variants that fall inside functional domains (B) or in an unsupervised manner by dividing the variants in groups with a high variant density (C). (partly adapted from Gelfman et al., 2019, p. 810)*

same domain can be of interest for future studies. On the other hand, Loehlein Fier et al. (2017) describe an unsupervised spatial clustering approach that splits all variants into groups where the distance between two consecutive variants on the genome is larger than expected (Figure 1C). This approach is based on the idea that disease-associated variants could lie closer together on the genome than random variants, for example because they cause a disruption of the same function of a gene (Loehlein Fier et al., 2017). Their method has already yielded promising results in data from Alzheimer's patients. While the spatial clustering method ensures that all the variants in the selected dataset are present in a cluster, one could argue that the clusters obtained from this method are overfitted to the set of variants that is used. Therefore, before general statements can be made about the coordinates of potential hotspots of ALS mutations, it is important to know how robust the clusters are that the spatial clustering algorithm finds. That is, only if a cluster is found in a certain part of the gene independently of the variant set that is used to assign the clusters, then we can make conclusions about potential hotpots found with this method.

The aim of this study is to research whether selecting variant sets using spatial clustering or functional domains can be used to replicate known hotspots of ALS mutations. If this technique replicates known hotspots, it is a promising technique to find new ALS hotspots and potentially even new ALS genes. We focus on three known ALS genes, where *SOD1* and *FUS* serve as a proof of concept and *NEK1* is analysed to apply the methods to a gene where it is unsure whether it contains hotspots of ALS mutations. A second aim of this study is to quantify the robustness of the clusters found using the spatial clustering method to establish if the results are specific to our dataset or not. That is, we aim to research whether the clusters we find based on our variant set are still found if the variant set that is used to assign clusters varies. The data analysis is done by extending an R package called RVAT (Rare-Variant Association Toolkit), designed specifically for research with rare variant association tests. RVAT brings together multiple components of RVB tests. Firstly, it allows for easy preparation of the data by building a database. In combination with a list of variants of interest, this database can be used for performing association tests. Additionally, lists of genes that should be tested can be combined with the results of association tests to run gene set analyses. Lastly, several methods are provided to intuitively visualise results with qqplots, manhattan plots, forest plots, and density plots. To facilitate data analysis of RVB tests, we have extended the package by building an interactive app with the R package Shiny.

Methods

Whole genome sequencing datasets

The first dataset with genomic variants used for this project is taken from Project MinE (Van Rheenen et al., 2016, 2018) and can be browsed on http://databrowser.projectmine.com. The ALS and control samples in this dataset have been aligned to genome build 37.75 and have been annotated using snpEFF (Single Nucleotide Polymorphism Effect) (Cingolani et al., 2012), as also described by Nicolas et al. (2018). This dataset contains 10,502 ALS samples and 26,035 controls. The b37.75 dataset has only been used to assess the robustness of the spatial clustering method. The second dataset used for this project is a currently unpublished extension of the original dataset aligned to genome build 38.105. This dataset contains 13,128 ALS samples and 69,775 controls. The b38.105 dataset has been used for RVB tests and calculating the robustness of the spatial clustering method. Quality control (QC) of the variants has been performed in accordance with the methods described by van Rheenen et al. (2021). This QC includes standard metrics such as correcting for population stratification and sex of the samples, in addition to the sequencing quality of the samples.

Mapping genomic coordinates to transcript coordinates

Both the spatial clustering method and the functional domains method were conducted with transcript coordinates. We chose to use transcript coordinates for the functional domains, because the coordinates of the domains we collected are also transcript-based. In the case of the spatial clustering, transcript coordinates were used because genomic coordinates introduce large distances between variants in different exons which would be absent in the protein. To map the genomic coordinates to transcript coordinates, we first selected the canonical transcript for each gene in the Ensembl b37.75 gtf or Ensembl b38.105 gtf (depending on the version of the dataset) by selecting the longest transcript based on total exon length. If multiple transcripts were longest, all were selected as canonical transcripts.

We distinguish between variants that fall inside the exons of a transcript and variants that flank the exons. For variants inside the exons, the transcript position was calculated by subtracting the cumulative width of the introns before the exons that the variant is in from the genomic position. Variants were said to flank an exon if they were within 12 base pairs on either side of the exon. These variants are included because they have a high probably of influencing splicing. The position of variants before an exon was changed to the first position of the exon; the positions of variants after an exon was changed to the last position of the exon.

Assignment of sets of variants

Supervised domain-based clustering

To make sets of variants found in functional domains, we retrieved the coordinates of Ensembl b38.105 domains for Interpro domains, coiled coils, transmembrane helices, low complexity regions, and cleavage sites. These were collected manually from the bioMart archive http://dec2021.archive.ensembl.org/biomart/martview/. Domain start and end coordinates were changed from amino acid positions to base pair positions to match the variant coordinates. Variant sets of all transcript-domain combinations in the dataset were made. If domains consisted of multiple unique ranges – either non-overlapping or overlapping but from different sources – each of these ranges was analysed as a separate domain. Transcript-domain combinations that spanned more than 90% of the width of the whole transcript were excluded from the analysis. The function *domainVarSet* from the RVAT package was used for assigning domain-based variant sets (Supplementary Materials 1). Separate sets of variants were made for variants previously classified as synonymous, moderate, or loss-of-function (LOF) variants. Synonymous variants are mutations that do not result in a change in amino acid. Moderate

variants do result in a change in amino acid, and for LOF variants this is predicted to disrupt the function of the protein.

Unsupervised spatial clustering

Our method for clustering variants along genes based on their genomic distance is based on the method described by Loehlein Fier et al. (2017). The genomic distance is defined as the number of base pairs between variants. In brief, this algorithm tries to find clusters of variants along a gene by comparing the median and mean distance between a group of consecutive variants. If the mean distance between all variants in a window exceeds the median distance, two variants in the group are said to be too far apart from each other to belong to the same cluster. This has been chosen as a metric because the mean is influenced strongly by outliers, whereas the median is not. Loehlein Fier et al. (2017) recommend choosing a group with 50-200 variants, i.e. a window size of 50 to 200. However, since the number of variants per gene in our gene set varies from 1 to 9539, it would neither be effective to choose a minimum window size of 50, nor to only choose one window size for all genes. Instead, we chose to assign larger window sizes as the number of variants in a gene increased. Specifically, for 0-15 variants the window size was 6, after which the window size was increased with 2 variants for each multiple of 16 variants (Equation 1).

$$window\ size = 6 + 2 * \left\lfloor \frac{number\ of\ variants}{16} \right\rfloor \qquad Equation\ 1$$

To ensure that distances between the last variant of a window and the first variant of the next window are not ignored, the windows overlap by a number of variants defined by the 'overlap' parameter. We selected the overlap to be half of the window-size so that each variant is selected twice. Using the function *genPartVarSet* in the RVAT package, for both the b37.75 and b38.105 dataset, we assigned sets of variants for each gene consisting of the clusters defined by the distances between the variants. All variants in the dataset, regardless of the phenotype of the samples or the MAF of the variants, were used to cluster the variants. Afterwards, separate sets of variants were made for moderate variants, synonymous variants, and LOF variants.

Robustness of the spatial clustering method

In order to estimate to what extent the results of the spatial clustering method can be generalised to other datasets, we need to know how robust the clusters are that are found. To estimate the robustness of the spatial clustering, we adapted the method to define the robustness of a clustering method described by Lu et al. (2019). In this method, the robustness of a clustering is found by repeating the clustering a large number of times with slightly different values for a selected parameter. In our case, this parameter is the number of samples in the dataset. Then, for each unique pair of units, e.g. samples or genes, we count in how many iterations this pair occurs in the same cluster. The robustness of this pair is defined as the number of co-occurrences (*o*) divided by the number of iterations (*r*). The robustness of the clustering method as a whole is the mean of the scores of all pairs. The number of pairs is indicated as *d* and the sum of all occurrences of all pairs together is *t*. This is summarised in Equation 2.

$$R = \frac{\sum o/r}{d} = \frac{t}{dr} \qquad Equation\ 2$$

We calculated the robustness for two types of units. Firstly, we took variants as unit. Since using a subset of the data can cause variants to disappear, the robustness of variant pairs containing these disappearing variants can artificially become low. We therefore adapted the formula by dividing the frequency of a variant pair by the number of iterations that the two

variants occurred in together, rather than by the total number of iterations. Secondly, we took the base positions of a transcript as units. Since base positions are always present in a transcript, regardless of the samples that are included in the dataset, we did not need to adjust the method in Equation 2. For each method, to compare the robustness of subsetted clustering with the clusters found using the whole dataset, we also calculated the robustness of the clustering using only the pairs of units that were present in the clustering based on the complete dataset. This shows whether the clusters that are found using the complete dataset are enriched compared to clusters that arise from taking subsets of the data.

The subsets of the data were made by selecting a certain number of samples and including all the variants that occurred at least once in any of the selected samples. We selected 500, 1000, 1500, 2500, 5000, 10,000, 15,000 and 20,000 random samples from both the b37.75 and b38.105 dataset. Additionally, we took all the ALS samples and an equally large number controls for both the b37.75 dataset (10,507 ALS samples and 10,507 controls) and the b38.105 dataset (13,128 ALS samples and 13,128 controls). Lastly, for the b38.105 dataset we took all controls and half of the ALS samples (6564 ALS samples and 69,775 controls). We repeated the subsetting 500 times for each number of samples. This process was performed for five ALS genes (*SOD1, FUS, TARDBP, NEK1,* and *C21orf2* or *CFAP410*), a potential ALS gene (*SCHBP1*) and four randomly selected genes (*FGR, VEGFA, NIPAL3,* and *SCYL3*).

Rare variant association testing

RVAT

The package RVAT facilitates each step of RVB tests, from making sets of variants to performing RVB tests and gene set analyses to visualising the results. An introduction to the package can be found at https://kkenna.github.io/rvat/. Most of the data infrastructure and functions necessary to perform association tests and visualise the results was already present before the start of this study. This data infrastructure is summarised in Supplementary Materials 2. To facilitate our analyses we extended the *rvatViewer* functionality and introduced the *genPartVarSet* and *domainVarSet* functionalities.

The *rvatViewer* is an app that facilitates interactive data analysis of RVB tests and gene set analyses. It has been built with the package R shiny. Using this *rvatViewer*, it is possible to make manhattan plots, qqplots, comparison plots, density plots, and forest plots on the appropriate object types. When coordinates of tracks such as domains and clusters are loaded, it is possible to zoom into the results of cluster/domain analyses or single-variant association test results per unit. This is illustrated in the following video: https://tinyurl.com/Unit-viewer-RVAT.

Burden tests

Burden tests were run on whole genes, clusters, and domains; the moderate, synonymous, and LOF sets of variants were all included. Three different combinations of filters were used: 1) minCarriers = 1, maxMAF = 0.05; 2) minCarriers = 1, maxMAF = 0.001; 3) minCarriers = 1, maxCarriers = 5. These filters reflect three types of variants: intermediate-frequency variants with a MAF between 0.001 and 0.05; rare variants with a MAF below 0.001; and ultra-rare variants with 1-5 carriers. The statistical test ‘firth’ was used, because this test resulted in acceptably low genomic inflation values. The genetic model ‘allelic’ was used, meaning that it is assumed that a heterozygous variant has less effect on a patient than a homozygous variant. We focussed on moderate variants only.

The results were filtered based on the P-values of the same units in control-control burden tests. That is, if a unit had a $P < 1.0x10^{-5}$ in the case-control burden tests, it was marked as a

potentially interesting result. A new burden test was run with all the potentially interesting sets of variants using only control samples to see if this result could be a false positive. For each control cohort, we tested all samples within that cohort against the samples that were not within that cohort. If a filter-unit-test combination had a P-value < $1.0x10^{-4}$ in the control-control burden test, this filter-unit-test combination was marked as a false positive result and removed from the results of the case-control burden tests. Not only does this remove false positive results, it also has the potential of reducing genomic inflation of the results. For zooming in on specific genes, five ALS genes were selected: *SOD1*, *FUS*, *NEK1, TARDBP* and *KIF5A* (the last two genes only as supplemental materials).

For the single variants that were studied from domains and clusters of interest, single-variant association results were generated using the statistical test 'firth'.

Domain set enrichment analysis

Gene set association tests were run on sets of domains. These sets of domains were defined as every occurrence of a domain across the genome. This resulted in a set of domains containing all coiled coils, one with all transmembrane domains, one with all low complexity regions, and one with all cleavage sites. Each Interpro domain was assigned its own domain set. We used a linear model as statistical test and a competitive null hypothesis, which means that we test for differential expression of domains in a domain set versus the other domains in the dataset. We also selected domain sets with <= 1000 domains, because at very large numbers of units the domain set becomes too biologically aspecific to give a meaningful interpretation of a potential significant result.

Results

To answer the question whether functional domains or spatial clusters of variants can be used to find hotspots of ALS mutations, various types of plots were used to visualise burden test results. We incorporated these visualisation methods in an app called the *rvatViewer* built with R Shiny. Specifically, to facilitate zooming in on a unit, be it a gene, functional domain, or spatial cluster, we built the 'Unit viewer' in this app.

Distinct patterns of association to ALS can be found across different ALS genes
We performed burden tests using the statistical test 'firth' on spatial clusters and domain-based clusters found in 13,128 patients and 69,775 controls. Both spatial clustering and domain-based clustering gave rise to clusters that recover known hotspots of ALS mutations. However, as can be seen in Figure 2, this is more often the case in variant sets defined by spatial clustering than by domain-based clustering. Figure 2 zooms in on two known ALS genes, namely *SOD1* and *FUS*. We see distinct patterns of associations in these two genes. In *SOD1*, strong associations with ALS are found across the whole gene, for both the domain-based clusters (orange blocks) and spatial clusters (blue blocks). In *FUS*, on the other hand, there is only one strong association with ALS found in a spatial cluster at the C-terminus of the gene. For the results of *TARDBP* and *KIF5A*, see Supplementary Materials 3.

The presence of a cluster or domain showing a significant association with ALS does not automatically imply that there is a hotspot of ALS mutations in that gene. Several sections of the gene could show moderately interesting results, making it difficult to say whether there is a heterogeneous distribution of causal variants at all and if so, which area of the gene is most interesting. We say there is an indication of the presence of a hotspot of ALS mutations in a gene if the P-value of the most significant domain or cluster, corrected for the multiple testing burden within the gene, is lower than the P-value of the whole gene. We need to correct for multiple testing within the gene, because we are testing the same variants multiple times, meaning that the results are not independent. The correction is done using the statistical method ACAT-O (Liu et al., 2019). This omnibus test combines the P-values of all domains and spatial clusters in a gene into one P-value for the whole gene while taking the sparsity of causal variants and potential differences in effect size of the domains and clusters in the gene

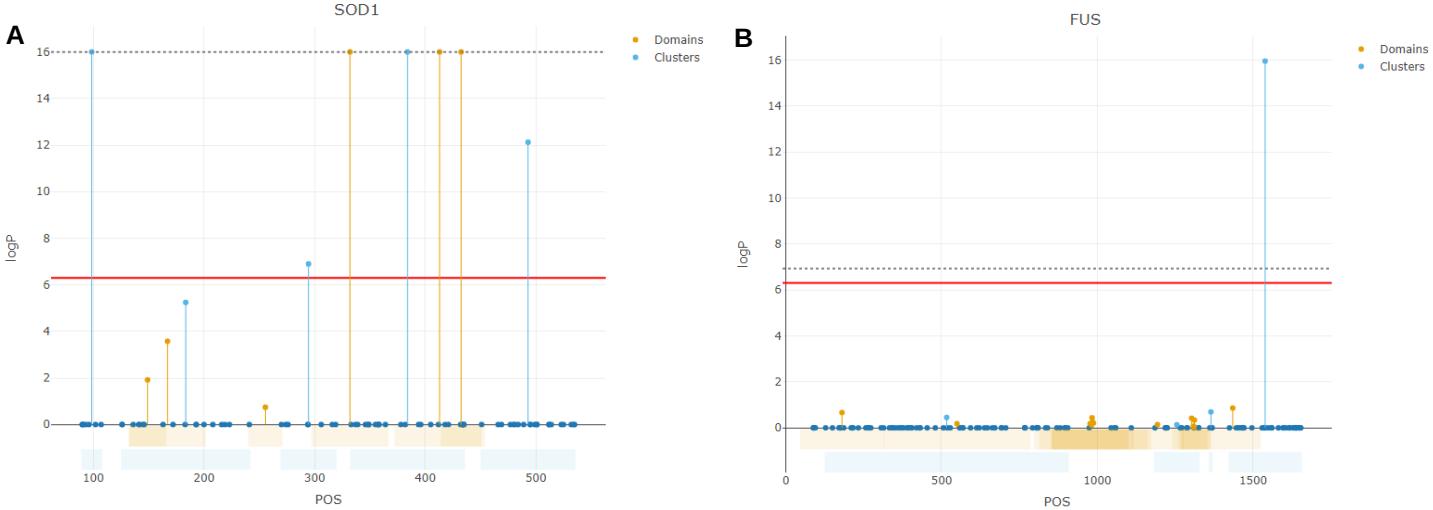


Figure 2 – Distinct patterns of ALS mutations can be found across ALS genes. *In these mutations plots, the blue dots are the locations of the variants inside a gene; the blocks underneath the x-axis indicate the functional domains (orange) and spatial clusters (blue). If the colour of the block is darker, there are overlapping domains in that region. The lollipops indicate the -log10(P)-value of RVB tests on the selected domains and clusters. The red line indicates the genome-wide significance threshold (around $0.05x10^{-6}$). The grey dotted line indicates the -log10(P)-value of an association test on the whole gene. P-values were cut at $1x10^{-16}$. A) There are no singular hotspots of ALS mutations in SOD1. B) There is a single hotspot of ALS mutations found in the C-terminus of FUS.*

into account. It is thus expected to give a P-value lower than the P-value of a burden test on the whole gene if a cluster of causal variants is present, and an equal or higher P-value if this is not the case.

In *SOD1* (Figure 2A), splitting up the gene in functional domains or spatial clusters gives rise to groups of variants throughout the whole gene that are either more significant (not visible in the figure, since P-values were cut at $1x10^{-16}$) or less significant than the whole gene. This is reflected in that the P-value of the whole gene and omnibus P-value of all domains and clusters are both smaller than $1x10^{-16}$. The fact that these P-values are essentially the same indicates that there is no single hotspot in *SOD1* that contains ALS-associated mutations.

In *FUS* (Figure 2B), the C-terminus is more significant in the spatial clusters than in the whole gene. However, this is not the case for the domain-based clustering results. Just like in *SOD1*, this is reflected in the P-values. *FUS* as a whole has a P-value of $1.17x10^{-7}$ (OR = 1.944). All the domains combined have an omnibus P-value of 0.498, which shows that the domains that were defined in *FUS* do not contain any rare variant signal. However, including the spatial clusters resulted in an omnibus P-value of $4.44x10^{-16}$, which is much lower than the P-value of the RVB test on the whole gene. The spatial cluster in question only contains rare variants. Four of these variants had more than 5 carriers and are therefore not considered ultra-rare. Two of those variants showed evidence of single variant signal ($P = 9.07x10^{-11}$ and $P = 2.99x10^{-7}$). However, the analysis of the cluster in question with only ultra-rare variants still results in a genome-wide significant signal ($P = 3.36x10^{-9}$, OR = 6.74). This implies that the signal seen in *FUS* is found in a hotspot of variants, rather than a single variant. These results are further supported by the fact that the effect size of the cluster in question is 6.644, more than three times as high as the effect size of the whole gene, indicating that the variants in this cluster are more important to ALS development than the other variants in the gene.

NEK1 contains both single-variant driven signal and an enrichment of rare damaging variants

NEK1 as a whole has a P-value of $1x10^{-16}$ (OR = 1.70) and an omnibus P-value of $1.48x10^{-14}$. In Figure 3, we see both significant results for functional domains and spatial clusters (for details on the variants in the selected clusters and domains, see Suppl. Materials 4). Firstly, the domain labelled as '3' is part of a low complexity region with a P-value of $3.87x10^{-14}$ and an effect size of 5.62. The removal of the rare variant in the cluster (rs199947197) increases the P-value to 0.75 (OR = 0.77), showing that this signal is driven by one of the variants in this domain, rather than the whole domain. Secondly, the spatial cluster labelled as '2' shows a significant location on the other side of *NEK1* ($P = 7.77x10^{-16}$, OR = 2.04). The removal of the intermediate-frequency variant (rs200161705) results in an increased P-value of $2.78x10^{-3}$ (OR

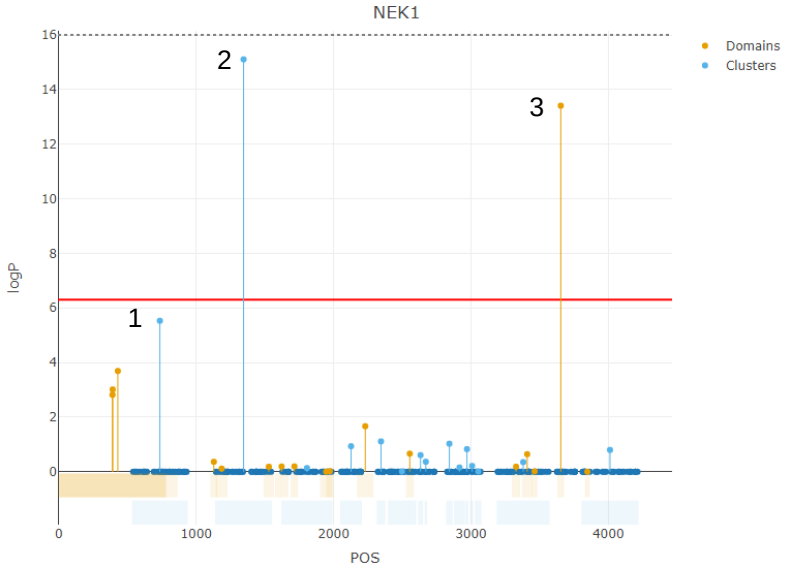


Figure 3 - NEK1 contains single-variant and rare variant-driven signal. *1) This set of variants at the N-terminus of NEK1 is driven by a set of rare and ultra-rare variants. 2) The signal in this cluster is mostly driven by a single, intermediate-frequency variant: rs200161705. Removal of this variant results in a P-value of $2.78x10^{-3}$. 3) The signal in this domain is mostly driven by a single rare variant: rs199947197. Removal of this variants results in a P-value of 0.75.*

= 1.98), which, while still more significant all the clusters and domain in the middle and end of the gene, indicates that a large part of the signal is driven by a single variant. Lastly, the cluster labelled '1', overlapping partly with three domains on the left, has a P-value of 2.95×10^{-6} (OR = 2.88) when including rare and ultra-rare variants. Including only ultra-rare variants gives a P-value of 2.48×10^{-4} (OR = 3.07). Combined with the fact that the P-value of all the regions that have not been discussed are much higher than those of clusters '1' and '2', we thus see an enrichment of rare causal variants in the N-terminus of *NEK1*. All in all, *NEK1* contains both single-variant driven signal and cluster-driven signal.

Prioritisation of candidate genes can be done using domain set analyses

A theoretical advantage of domain-based clustering is that if a domain is significant in one gene, it could also be an interesting target in another gene. Because domains are annotated according to their function, it is possible to test the association with ALS of all occurrences of a functional domain across the genome. If such a domain set has a significant association with ALS while none of the separate domains are significant, this could point at cellular function that is important in, for example, ALS pathology. However, domain sets can only be an interesting target if they do not contain a very large number of genes. This makes the domain too aspecific to be biologically relevant.

As an example of this use of domain-based variant sets, we defined a domain set as all the occurrences of a domain across the genome. The functional domains from Ensembl are thus divided into domain sets consisting of all low-complexity domains, all transmembrane helices, all coiled coils, all cleavage sites, or Interpro domains with a specific function. To calculate the significance of the association between that domain set and ALS, we then compare the burden test results of the domains in the domain set with those outside of the domain set. When filtering for biologically specific domain sets (<1000 occurrences of a domain across the genome), we are left with only Interpro domains (Table 1). The three most significant of these results are domain sets that contain a domain present in *SOD1*. If the results obtained from using single domains as domain sets are promising, the analysis could be extended to make domain sets of, for example, domain families.

Preliminary evaluations of the robustness of the spatial clustering method

As opposed to domain-based clustering, spatial clustering is completely tailored to the selected dataset. The inclusion or exclusion of samples can change the variant pool in such a way that different clusters are found than with another set of samples, because the distances between the variants change. This means that the clusters obtained with the spatial clustering method

geneSetName	Function	Ngenes	P	Effect
IPR024134	Superoxide dismutase (Cu/Zn) / superoxide dismutase copper chaperone	3	7.92×10^{-6}	2.693
IPR001424	Superoxide dismutase, copper/zinc binding domain	2	1.16×10^{-5}	3.233
IPR036423	Superoxide dismutase-like, copper/zinc binding domain superfamily	3	8.78×10^{-5}	2.341
IPR026261	RanBP-type and C3HC4-type zinc finger-containing protein 1/SHARPIN	2	1.39×10^{-4}	2.778
IPR043197	Plakin	5	4.30×10^{-4}	1.611

Table 1 – The top results of domain set association tests reveal previously found protein functions related to ALS. *The three most significant results, whose functions are related the oxidation pathways of the cell, were driven by SOD1.*

can easily be overfitted to the data. To examine this possibility, we did a preliminary evaluation of the robustness of the spatial clusters. We used a method adapted from those described by Lu et al. (2019). In this method, one aspect of the clustering is changed across a number of iterations, after which the mean frequency is taken of the occurrence of each pair of units in the same cluster across all iterations. The more frequently pairs of units co-occur, the more often the clusters are identical. Thus, we can say that the more frequently pairs of units co-occur, the more robust the spatial clustering method is. The parameter we changed was the number of samples that is included in the dataset. We used two different types of unit to calculate the robustness of the spatial clustering method: variants and base positions.

The rationale behind using variants as units is that the positions of the variants is what determines the location of the cluster. If we compare which variants make up each cluster across iterations, we thus know how the clusters change with differing sample sizes (Supplementary Materials 5). A drawback of this approach is that if a variant in the middle of a cluster is absent in some subsets of the samples, the score of that cluster decreases, even though the cluster still covers the same area on the gene. Since we are looking at spatial clusters, and not just groups of units, one could argue that the base positions that the clusters cover are more important than the variants present in the clusters. Thus, by using base positions as units, we can calculate how robustly the clusters cover certain areas of the gene. An additional advantage to using base positions as units is that they are not unique to our dataset, as opposed to dataset-specific variants, so we can better generalise the robustness of the method to other datasets.

Generally speaking, Figure 4 (blue datapoints) shows that lowering the number of samples adversely affects the robustness of the clustering. This relates to the fact that many of the variants in our dataset have few carriers, so the more samples are left out, the more likely it is that variants will not be sampled. Therefore, the distances between the remaining variants can change such that new groups of variants emerge. If clusters change, their position changes, which reduces their robustness. Vice versa, the more samples are included, the more dense potential groups of variants will be and the higher the likelihood that variants are frequently clustered together.

The method described above calculates the robustness of clusters relative to the clusters found across the iterations, not in relation to a reference clustering. Because we want to know to what extent the clusters we found in the whole dataset are overfitted to our data or not, we also attempted to calculate the robustness of the clustering of a subset of the data compared to the clustering based on the whole dataset. We did this by comparing the robustness of the position pairs that are present in the original clustering to the robustness of all the position pairs (Figure 4; yellow datapoints). This showed that the positions that co-occur in the original cluster also co-occur more often across all iterations than the position pairs that arise due to changes in the clustering. In turn, we can thus say that the clusters found using the whole dataset are more robust than other potential clusters, regardless of the sample size.

While the maximum robustness achieved using the method calculating robustness relatively across iterations is only 0.474 out of a possible 1.0, the fact that clusters present in the original clustering have a higher robustness (0.799) supports that the spatial clustering method can relatively reliably find similar clusters, regardless of the samples in the dataset. Additional support for this is that the spatial clustering method gave rise to clusters that confirm previous findings of ALS hotspots (Figure 2).

Discussion

The aim of this study was to research whether selecting variant sets using spatial clustering or domain-based clustering can be used to find hotspots of ALS-associated mutations in known

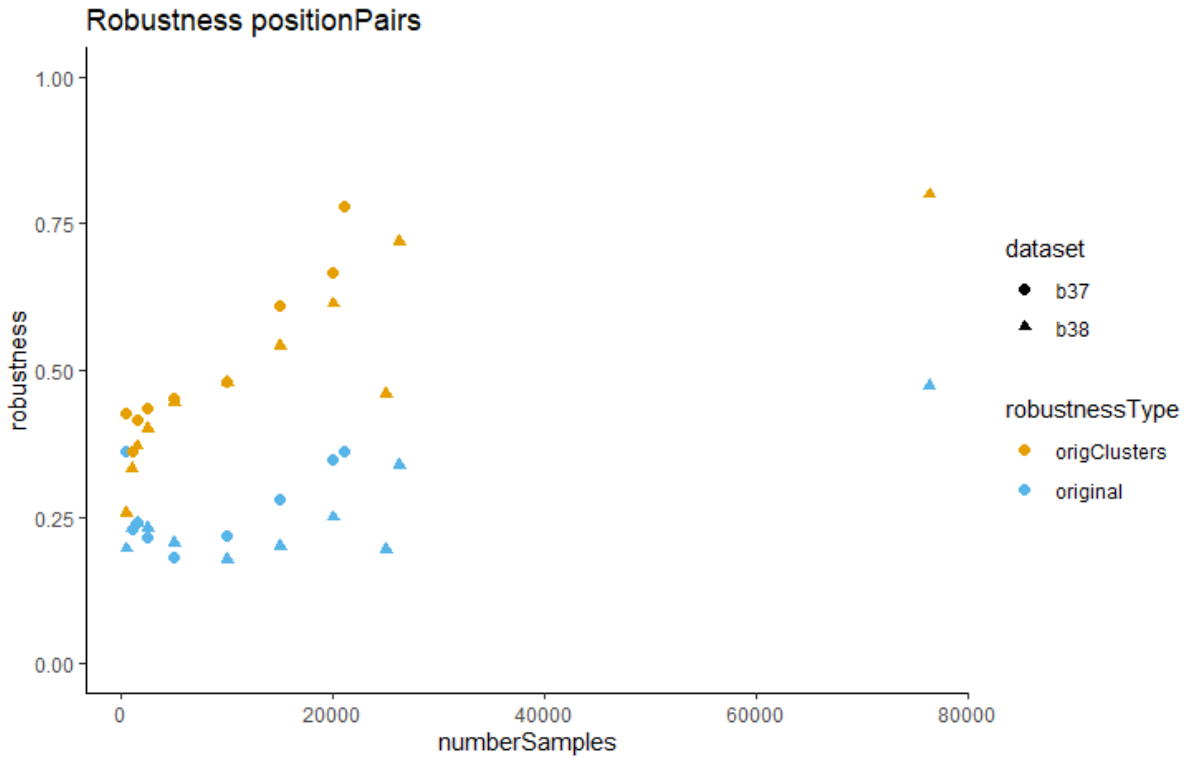


Figure 4 - Robustness of spatial clusters based on the complete dataset is higher than that of all clusters found by subsetting the dataset. *For the b37.75 and b38.105 dataset, several subsets of samples were taken 500 times per sample size. For each of these subsets, the robustness of all base position pairs was calculated. The blue data points (bottom row) indicates the robustness calculated with the original method described by Lu et al., (2019). The golden data points (top row) indicates the robustness scores calculated when only using the position pairs that occurred in the clustering of the whole dataset. The difference in score implies that the position pairs found in the original clustering, and by extension the clusters in the original clustering, are more robust than the other possible clusters that could be found.*

ALS genes. If this is the case, these methods could potentially also be used to search for new hotspots or new genes associated with ALS. For the clustering methods to be successful, they need to recover known ALS hotspots after filtering out common variants.

For this exploratory analysis, we focussed on *FUS*, *SOD1*, and *NEK1*. In the case of *FUS*, we found a significant cluster in the C-terminus of the gene using the spatial clustering method. For *SOD1*, we found several strongly associated regions spread out over the whole gene with both clustering methods. In the case of *NEK1*, for the domain-based clustering method we found a cluster that was entirely driven by a single rare variant. Using the spatial clustering method, we found one significant variant cluster largely driven by a single intermediate-frequency variant, and a cluster at the N-terminus that was driven by the complete set of rare and ultra-rare variants in that cluster. All of these clusters had a positive effect size, which is to be expected of genes that are known to be related to an aspect of ALS.

Interpretation of inspected genes

Our results show that of the three inspected genes, the spatial clustering methods resulted in significant results that could potentially point at hotspots of ALS mutations in *SOD1*, *FUS*, and *NEK1*. The domain-based clustering method, however, did not find potential hotspots for *FUS*. Firstly, *SOD1* has not been shown to have hotspots of ALS variants; rather, the causal variants are spread throughout the whole gene (shown for example in figure 1 of Ruffo et al., 2022). Thus, the large number and spread of significant functional domains and spatial clusters that we found throughout *SOD1* supports previous findings.

Secondly, *FUS* has been shown to contain hotspots of ALS mutations in the C-terminus of the gene (Kwiatkowski et al., 2009; Shang & Huang, 2016). According to Shang & Huang

(2016), this hotspot is located at an RGG-repeat region and especially the NLS, ranging from amino acids 453-501 and 510-526 respectively. Figure 2B shows that the NLS is not covered by any of the functional domains in our dataset, explaining why we did not find a domain-based signal in *FUS*. The spatial clustering method, on the other hand, did give rise to a cluster that covered the variants in and around the NLS. The cluster contains two rare single variants that are significantly associated with ALS. Removing these variants from the cluster resulted in a higher, but still genome-wide significant, P-value for the cluster. This implies that even though the presence of the single variants makes the cluster more significant, the signal is still spread over the whole cluster, rather than a single variant. We can therefore conclude that there is indeed a hotspot of ALS mutations in the C-terminus of *FUS* as suggested by Shang & Huang (2016).

Lastly, there is no evidence yet regarding ALS hotspots in *NEK1*. We found three interesting areas in this gene. Using the domain-based clustering method, we found a genome-wide significant domain containing four ultra-rare variants (1-5 carriers) and one rare variant (MAF < 0.001). However, removal of the rare variant from the cluster resulted in a non-significant result (P = 0.75, OR = 0.77), indicating that the signal in this cluster was caused by the single LOF variant rs199947197, rather than the whole cluster. This variant is located near the C-terminus of *NEK1*, specifically around a nuclear export signal (Nguyen et al., 2018) and a coiled coil segment that has been reported to be involved in the interaction of *NEK1* with the ATR-ATRIP complex (Melo-Hanchuk et al., 2017). This interaction is important in a *NEK1*-mediated DNA damage response by *ATR* (S. Liu et al., 2013), meaning that a mutation in this coiled coil domain could result in a reduced DNA damage response. In addition to the domain-based cluster, we found a genome-wide significant spatial cluster closer to the N-terminus of *NEK1*. Similar to the domain-based cluster, removal of the most common variant in this cluster causes the P-value to increase to 2.78×10^{-3} (OR = 1.98). While this is not genome-wide significant, it is still more significant than all the other domains and clusters towards the C-terminus of *NEK1*. Lastly, at the N-terminus of *NEK1* we found a cluster located in the kinase domain of *NEK1* that showed a stronger association of rare and ultra-rare variants with ALS than the spatial cluster next to it. The signal in this area of the gene can therefore be associated with the aforementioned involvement of *NEK1* in a DNA-damage response (S. Liu et al., 2013). To conclude, apart from the LOF variant rs199947197, most of the signal in *NEK1* is found in an enrichment of variants at the N-terminus of the gene, which is important in the function of *NEK1* in a DNA-damage response.

Evaluation of domain-based clustering and spatial clustering

Both the domain-based clustering method and the spatial clustering method have strengths and limitations. The first strength of the domain-based clustering method is that the domains are already functionally annotated, meaning that if we find a significant association between a domain and ALS, these annotations can aid in the interpretation of the results. The second strength of this method is that the coordinates we obtain for the domains have been defined independently of our dataset, which makes it possible to generalise potential results to other datasets. However, the fact that we are dependent on other studies for the coordinates of the domains also points at a limitation of the domain-based method. The specific dataset with domain coordinates that is selected for the analysis can influence the final results that we get. Not only might domains not be available for all genes or transcripts in the genome, domains do not always cover the entire gene. While this is to be expected of functional domains, it also means there is a chance that mutational hotspots are missed. This limitation could partially be solved by selecting a different set of domain coordinates. For example, Gelfman et al. (2019) perform domain-based clustering based on functional domains collected and described by Gussow et al. (2016) and Marchler-Bauer et al. (2013). Gelfman et al. (2019) define a functional domain as either a domain in the Conserved Domain Database (CCD) (Marchler-

Bauer et al., 2013) or the spaces between two of those domains in the same gene. This approach solves the problem of potentially missing important variants. However, with regards to *FUS*, the only significant cluster that they found was actually a group of variants between CCD domains, which overlapped with the aforementioned RGG-repeat. Essentially, this means that they still did not find a functional domain in *FUS* present in their dataset that was associated with ALS, so the problem of being dependent on the selected dataset with domain coordinates remains.

The aforementioned weakness of domain-based clustering is absent from the spatial clustering method. Clusters defined by spatial clustering cover the whole gene and do not miss signal at the end of the gene, because these variants are also included in the clusters. However, one could argue that the clusters that are found using the spatial clustering method are overfitted to the dataset, since the clusters are based on the distance between only the variants included in the selected dataset. Our exploratory analysis of the robustness of this method (based on Lu et al., 2019), however, implies that the clusters that are found are not completely dependent on the dataset, but occur more frequently than other potential spatial clusters. To further study the robustness of the spatial clusters across datasets, we could use the spatial clustering algorithm on other datasets, such as the data provided by the Genome Aggregation Database (gnomAD) (Karczewski et al., 2020). An additional point to keep in mind regarding the spatial clustering of variants is that the spatial clustering method does not have the advantage of the domain-based clustering method that a function has already been described for significant results. Spatial clusters are purely based on the distances between variant positions in a transcript, not on biological functions. In order to find a potential biological function behind a significant cluster, it is thus still necessary to look at the domain structure of a protein.

To conclude this section on spatial clustering, we want to point out that the method could also be used for different purposes than just finding groups of variants. For example, the method could be used to find differences in variant positions between cases and controls. An R package called DoEstRare has been developed that looks for clusters of disease-associated variants and evaluates differences in variant positions between cases and controls (Persyn et al., 2017). Both the clustering method described by Loehlein Fier et al. (2017) and the method to find differences in gene structure between cases and controls described by Persyn et al. (2017) have been tested on Alzheimer's and not ALS. However, the fact that we found evidence of hotspots of ALS mutations using spatial clustering suggests that we might find differences in gene structure between cases and controls in ALS as well.

Whether the domain-based clustering or the spatial clustering method is most suitable for a specific research question depends on the aim of the research. If the aim is to research a process or cell function associated with ALS, the best approach would be domain-based clustering, since functional domains are annotated so that domains with the function of interest can be selected for analysis. Alternatively, the annotations can be used to find new processes related to ALS. This can be done by analysing single domains, domain sets of single domains, or domain sets of domain families. An example of a method that tries to find signals in domain sets of single domains is the R package REBET (Zhu et al., 2018). This package first defines groups of variants based on function and functional impact, after which it tries all combinations of clusters within a gene to find the selection of clusters that collectively show the strongest association to a disease.

If the aim is to find single rare variants that could be the cause of part of the disease phenotype seen in a sample, then the spatial clustering method would be better suited. Considering a traditional RVB test takes all variants in a gene and gives one output, it can be hard to find variants in significantly associated genes that are an interesting target for further research. By using the spatial clustering method, you narrow down the scope of potentially

interesting variants to only the variants in a significant cluster, rather than all the variants in the entire gene. This facilitates the selection of variants that cause the ALS phenotype and thus furthers our knowledge of the genetic component of ALS.

In the end, the best results are likely obtained using a combination of both methods, since neither focussing on the position nor the function of a variant cluster is guaranteed to capture all promising variant groupings. Combining the results of these methods is especially powerful when significant clusters and domains (partially) overlap. This can simultaneously facilitate narrowing down on the potentially interesting variants in a cluster and finding the function of the protein that is disrupted by these variants.

Impact and future research

The use of a clustering method for association testing is important to study which area of associated genes actually cause the signal of the whole gene. While rare-variant association tests such as SKAT (S. Lee et al., 2014) and ACAT-V (Y. Liu et al., 2019) are designed to find genes in which smaller numbers of variants are causal amidst variants with potentially opposing effects, these tests do not identify in which region of the gene the smaller groups of causal variants are located. Thus, the clustering methods can help us narrow down which areas of the genes are important, even if this signal is not driven by a single intermediate-frequency variant, but a group of rare or ultra-rare variants.

Future research into the topic of RVB tests on spatial clusters can focus on the quantifying the robustness of the spatial clusters found using our dataset. As previously mentioned, one approach for this could be to collect data from other datasets, such as gnomAD (Karczewski et al., 2020), and comparing the locations of the clusters found in those datasets to the clusters we have found in this study. Additionally, future studies can look into how a spatial cluster is actually defined in space. Currently the method divides variants in groups based on one-dimensional differences between them, i.e. the distance in base-pairs on a transcript. However, not all variants that are far apart in a transcript are also far apart in the protein. Therefore, it would be worth researching whether clusters of variants can be found in three-dimensional models of proteins that have a stronger association with ALS than other regions of the gene.

Sources

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, *6*(2), 80–92. https://doi.org/10.4161/fly.19695

Cirulli, E. T., White, S., Read, R. W., Elhanan, G., Metcalf, W. J., Tanudjaja, F., Fath, D. M., Sandoval, E., Isaksson, M., Schlauch, K. A., Grzymski, J. J., Lu, J. T., & Washington, N. L. (2020). Genome-wide rare variant analysis for thousands of phenotypes in over 70,000 exomes from two cohorts. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-14288-y

Cooper-Knock, J., Moll, T., Ramesh, T., Castelli, L., Beer, A., Robins, H., Fox, I., Niedermoser, I., Van Damme, P., Moisse, M., Robberecht, W., Hardiman, O., Panades, M. P., Assialioui, A., Mora, J. S., Basak, A. N., Morrison, K. E., Shaw, C. E., Al-Chalabi, A., … Shaw, P. J. (2019). Mutations in the Glycosyltransferase Domain of GLT8D1 Are Associated with Familial Amyotrophic Lateral Sclerosis. *Cell Reports*, *26*(9), 2298-2306.e5. https://doi.org/10.1016/j.celrep.2019.02.006

DeJesus-Hernandez, M., Mackenzie, I. R., Boeve, B. F., Boxer, A. L., Baker, M., Rutherford, N. J., Nicholson, A. M., Finch, N. C. A., Flynn, H., Adamson, J., Kouri, N., Wojtas, A., Sengdy, P., Hsiung, G. Y. R., Karydas, A., Seeley, W. W., Josephs, K. A., Coppola, G., Geschwind, D. H., … Rademakers, R. (2011). Expanded GGGGCC Hexanucleotide Repeat in Noncoding Region of C9ORF72 Causes Chromosome 9p-Linked FTD and ALS. *Neuron*, *72*(2), 245–256. https://doi.org/10.1016/j.neuron.2011.09.011

Ferraiuolo, L., Meyer, K., Sherwood, T. W., Vick, J., Likhite, S., Frakes, A., Miranda, C. J., Braun, L., Heath, P. R., Pineda, R., Beattie, C. E., Shaw, P. J., Askwith, C. C., McTigue, D., & Kaspar, B. K. (2016). Oligodendrocytes contribute to motor neuron death in ALS via SOD1-dependent mechanism. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(42), E6496–E6505. https://doi.org/10.1073/pnas.1607496113

Gelfman, S., Dugger, S., de Araujo Martins Moreno, C., Ren, Z., Wolock, C. J., Shneider, N. A., Phatnani, H., Cirulli, E. T., Lasseigne, B. N., Harris, T., Maniatis, T., Rouleau, G. A., Brown, R. H., Gitler, A. D., Myers, R. M., Petrovski, S., Allen, A., Goldstein, D. B., & Harms, M. B. (2019). A new approach for rare variation collapsing on functional protein domains implicates specific genic regions in ALS. *Genome Research*, *29*(5), 809–818. https://doi.org/10.1101/gr.243592.118

Gussow, A. B., Petrovski, S., Wang, Q., Allen, A. S., & Goldstein, D. B. (2016). The intolerance to functional genetic variation of protein domains predicts the localization of pathogenic mutations within genes. *Genome Biology*, *17*(1), 1–11. https://doi.org/10.1186/s13059-016-0869-4

Hardiman, O., Al-Chalabi, A., Chio, A., Corr, E. M., Logroscino, G., Robberecht, W., Shaw, P. J., Simmons, Z., & Van Den Berg, L. H. (2017). Amyotrophic lateral sclerosis. *Nature Reviews Disease Primers*, *3*. https://doi.org/10.1038/nrdp.2017.71

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A., … Daly, M. J. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, *581*(7809), 434–443. https://doi.org/10.1038/s41586-020-2308-7

Kenna, K. P., Van Doormaal, P. T. C., Dekker, A. M., Ticozzi, N., Kenna, B. J., Diekstra, F.

P., Van Rheenen, W., Van Eijk, K. R., Jones, A. R., Keagle, P., Shatunov, A., Sproviero, W., Smith, B. N., Van Es, M. A., Topp, S. D., Kenna, A., Miller, J. W., Fallini, C., Tiloca, C., … Castellotti, B. (2016). NEK1 variants confer susceptibility to amyotrophic lateral sclerosis. *Nature Genetics*, *48*(9), 1037–1042. https://doi.org/10.1038/ng.3626

Kwiatkowski, T. J., Bosco, D. A., LeClerc, A. L., Tamrazian, E., Vanderburg, C. R., Russ, C., Davis, A., Gilchrist, J., Kasarskis, E. J., Munsat, T., Valdmanis, P., Rouleau, G. A., Hosler, B. A., Cortelli, P., De Jong, P. J., Yoshinaga, Y., Haines, J. L., Pericak-Vance, M. A., Yan, J., … Brown, R. H. (2009). Mutations in the FUS/TLS gene on chromosome 16 cause familial amyotrophic lateral sclerosis. *Science*, *323*(5918), 1205–1208. https://doi.org/10.1126/science.1166066

Lattante, S., Marangi, G., Doronzio, P. N., Conte, A., Bisogni, G., Zollino, M., & Sabatelli, M. (2020). High-throughput genetic testing in ALS: The challenging path of variant classification considering the acmg guidelines. *Genes*, *11*(10), 1–31. https://doi.org/10.3390/genes11101123

Lee, Seunggeun, Wu, M. C., & Lin, X. (2012). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, *13*(4), 762–775. https://doi.org/10.1093/biostatistics/kxs014

Lee, Seunggeung, Abecasis, G. R., Boehnke, M., & Lin, X. (2014). Rare-variant association analysis: Study designs and statistical tests. *American Journal of Human Genetics*, *95*(1), 5–23. https://doi.org/10.1016/j.ajhg.2014.06.009

Liu, S., Ho, C. K., Ouyang, J., & Zou, L. (2013). Nek1 kinase associates with ATR-ATRIP and primes ATR for efficient DNA damage signaling. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(6), 2175–2180. https://doi.org/10.1073/pnas.1217781110

Liu, Y., Chen, S., Li, Z., Morrison, A. C., Boerwinkle, E., & Lin, X. (2019). ACAT: A Fast and Powerful p Value Combination Method for Rare-Variant Analysis in Sequencing Studies. *American Journal of Human Genetics*, *104*(3), 410–421. https://doi.org/10.1016/j.ajhg.2019.01.002

Loehlein Fier, H., Prokopenko, D., Hecker, J., Cho, M. H., Silverman, E. K., Weiss, S. T., Tanzi, R. E., & Lange, C. (2017). On the association analysis of genome-sequencing data: A spatial clustering approach for partitioning the entire genome into nonoverlapping windows. *Genetic Epidemiology*, *41*(4), 332–340. https://doi.org/10.1002/gepi.22040

Lu, Y., Phillips, C. A., & Langston, M. A. (2019). A robustness metric for biological data clustering algorithms. *BMC Bioinformatics*, *20*(Suppl 15), 1–8. https://doi.org/10.1186/s12859-019-3089-6

Majounie, E., Renton, A. E., Mok, K., Dopper, E. G. P., Waite, A., Rollinson, S., Chiò, A., Restagno, G., Nicolaou, N., Simon-Sanchez, J., van Swieten, J. C., Abramzon, Y., Johnson, J. O., Sendtner, M., Pamphlett, R., Orrell, R. W., Mead, S., Sidle, K. C., Houlden, H., … Logroscino, G. (2012). Frequency of the C9orf72 hexanucleotide repeat expansion in patients with amyotrophic lateral sclerosis and frontotemporal dementia: A cross-sectional study. *The Lancet Neurology*, *11*(4), 323–330. https://doi.org/10.1016/S1474-4422(12)70043-1

Marchler-Bauer, A., Zheng, C., Chitsaz, F., Derbyshire, M. K., Geer, L. Y., Geer, R. C., Gonzales, N. R., Gwadz, M., Hurwitz, D. I., Lanczycki, C. J., Lu, F., Lu, S., Marchler, G. H., Song, J. S., Thanki, N., Yamashita, R. A., Zhang, D., & Bryant, S. H. (2013). CDD: Conserved domains and protein three-dimensional structure. *Nucleic Acids Research*, *41*(D1), 348–352. https://doi.org/10.1093/nar/gks1243

Melo-Hanchuk, T. D., Slepicka, P. F., Meirelles, G. V., Basei, F. L., Lovato, D. V., Granato, D. C., Pauletti, B. A., Domingues, R. R., Leme, A. F. P., Pelegrini, A. L., Lenz, G., Knapp, S., Elkins, J. M., & Kobarg, J. (2017). NEK1 kinase domain structure and its dynamic protein interactome after exposure to Cisplatin. *Scientific Reports*, *7*(1), 1–13. https://doi.org/10.1038/s41598-017-05325-w

Nguyen, H. P., Van Mossevelde, S., Dillen, L., De Bleecker, J. L., Moisse, M., Van Damme, P., Van Broeckhoven, C., van der Zee, J., Engelborghs, S., Crols, R., De Deyn, P. P., De Jonghe, P., Baets, J., Cras, P., Mercelis, R., Vandenberghe, R., Sieben, A., Santens, P., Ivanoiu, A., … Delbeck, J. (2018). NEK1 genetic variability in a Belgian cohort of ALS and ALS-FTD patients. *Neurobiology of Aging*, *61*, 255.e1-255.e7. https://doi.org/10.1016/j.neurobiolaging.2017.08.021

Nicolas, A., Kenna, K., Renton, A. E., Ticozzi, N., Faghri, F., Chia, R., Dominov, J. A., Kenna, B. J., Nalls, M. A., Keagle, P., Rivera, A. M., van Rheenen, W., Murphy, N. A., van Vugt, J. J. F. A., Geiger, J. T., van der Spek, R., Pliner, H. A., Shankaracharya, Smith, B. N., … Traynor, B. J. (2018). Genome-wide Analyses Identify KIF5A as a Novel ALS Gene. *Neuron*, *97*(6), 1268-1283.e6. https://doi.org/10.1016/j.neuron.2018.02.027

Pal, S., Tiwari, A., Sharma, K., & Sharma, S. K. (2020). Does conserved domain SOD1 mutation has any role in ALS severity and therapeutic outcome? *BMC Neuroscience*, *21*(1), 1–17. https://doi.org/10.1186/s12868-020-00591-3

Persyn, E., Karakachoff, M., Le Scouarnec, S., Le Clézio, C., Campion, D., Schott, J. J., Redon, R., Bellanger, L., Dina, C., Génin, E., Campion, D., Dartigues, J. F., Deleuze, J. F., Lambert, J. C., Ludwig, T., Grenier-Boley, B., Letort, S., Lindenbaum, P., Meyer, V., … Lechner, D. (2017). DoEstRare: A statistical test to identify local enrichments in rare genomic variants associated with disease. *PLoS ONE*, *12*(7), 1–20. https://doi.org/10.1371/journal.pone.0179364

Povysil, G., Petrovski, S., Hostyk, J., Aggarwal, V., Allen, A. S., & Goldstein, D. B. (2019). Rare-variant collapsing analyses for complex traits: guidelines and applications. *Nature Reviews Genetics*, *20*(12), 747–759. https://doi.org/10.1038/s41576-019-0177-4

Ruffo, P., Perrone, B., & Conforti, F. L. (2022). SOD-1 Variants in Amyotrophic Lateral Sclerosis: Systematic Re-Evaluation According to ACMG-AMP Guidelines. *Genes*, *13*(3). https://doi.org/10.3390/genes13030537

Shang, Y., & Huang, E. J. (2016). Mechanisms of FUS mutations in familial amyotrophic lateral sclerosis. *Brain Research*, *1647*, 65–78. https://doi.org/10.1016/j.brainres.2016.03.036

Suk, T. R., & Rousseaux, M. W. C. (2020). The role of TDP-43 mislocalization in amyotrophic lateral sclerosis. *Molecular Neurodegeneration*, *15*(1), 1–16. https://doi.org/10.1186/s13024-020-00397-1

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., & Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, *20*(8), 467–484. https://doi.org/10.1038/s41576-019-0127-1

Thomas, D. C., Haile, R. W., & Duggan, D. (2005). Recent developments in genomewide association scans: A workshop summary and review. *American Journal of Human Genetics*, *77*(3), 337–345. https://doi.org/10.1086/432962

Urushitani, M., Kurisu, J., Tsukita, K., & Takahashi, R. (2002). Proteasomal inhibition by misfolded mutant superoxide dismutase 1 induces selective motor neuron death in familial amyotrophic lateral sclerosis. *Journal of Neurochemistry*, *83*(5), 1030–1042. https://doi.org/10.1046/j.1471-4159.2002.01211.x

Van Rheenen, W., Pulit, S. L., Dekker, A. M., Al Khleifat, A., Brands, W. J., Iacoangeli, A., Kenna, K. P., Kavak, E., Kooyman, M., McLaughlin, R. L., Middelkoop, B., Moisse, M., Schellevis, R. D., Shatunov, A., Sproviero, W., Tazelaar, G. H. P., Van der Spek, R. A. A., Van Doormaal, P. T. C., Van Eijk, K. R., … Veldink, J. H. (2018). Project MinE: study design and pilot analyses of a large-scale whole-genome sequencing study in amyotrophic lateral sclerosis. *European Journal of Human Genetics*, *26*(10), 1537–1546. https://doi.org/10.1038/s41431-018-0177-4

Van Rheenen, W., Shatunov, A., Dekker, A. M., McLaughlin, R. L., Diekstra, F. P., Pulit, S. L., Van Der Spek, R. A. A., Võsa, U., De Jong, S., Robinson, M. R., Yang, J., Fogh, I., Van Doormaal, P. T. C., Tazelaar, G. H. P., Koppers, M., Blokhuis, A. M., Sproviero, W., Jones, A. R., Kenna, K. P., … Veldink, J. H. (2016). Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. *Nature Genetics*, *48*(9), 1043–1048. https://doi.org/10.1038/ng.3622

van Rheenen, W., van der Spek, R. A. A., Bakker, M. K., van Vugt, J. J. F. A., Hop, P. J., Zwamborn, R. A. J., de Klein, N., Westra, H. J., Bakker, O. B., Deelen, P., Shireby, G., Hannon, E., Moisse, M., Baird, D., Restuadi, R., Dolzhenko, E., Dekker, A. M., Gawor, K., Westeneng, H. J., … Veldink, J. H. (2021). Common and rare variant association analyses in amyotrophic lateral sclerosis identify 15 risk loci with distinct genetic architectures and neuron-specific biology. *Nature Genetics*, *53*(12), 1636–1648. https://doi.org/10.1038/s41588-021-00973-1

Vasta, R., Chia, R., Traynor, B. J., & Chiò, A. (2022). Unraveling the complex interplay between genes, environment, and climate in ALS. *EBioMedicine*, *75*, 1–12. https://doi.org/10.1016/j.ebiom.2021.103795

Zhu, B., Mirabello, L., & Chatterjee, N. (2018). A Subregion-based Burden Test for Simultaneous Identification of Susceptibility oci and Sub-regions within. *Genetic Epidemiology*, *42*(7), 673–683. https://doi.org/10.1002/gepi.22134

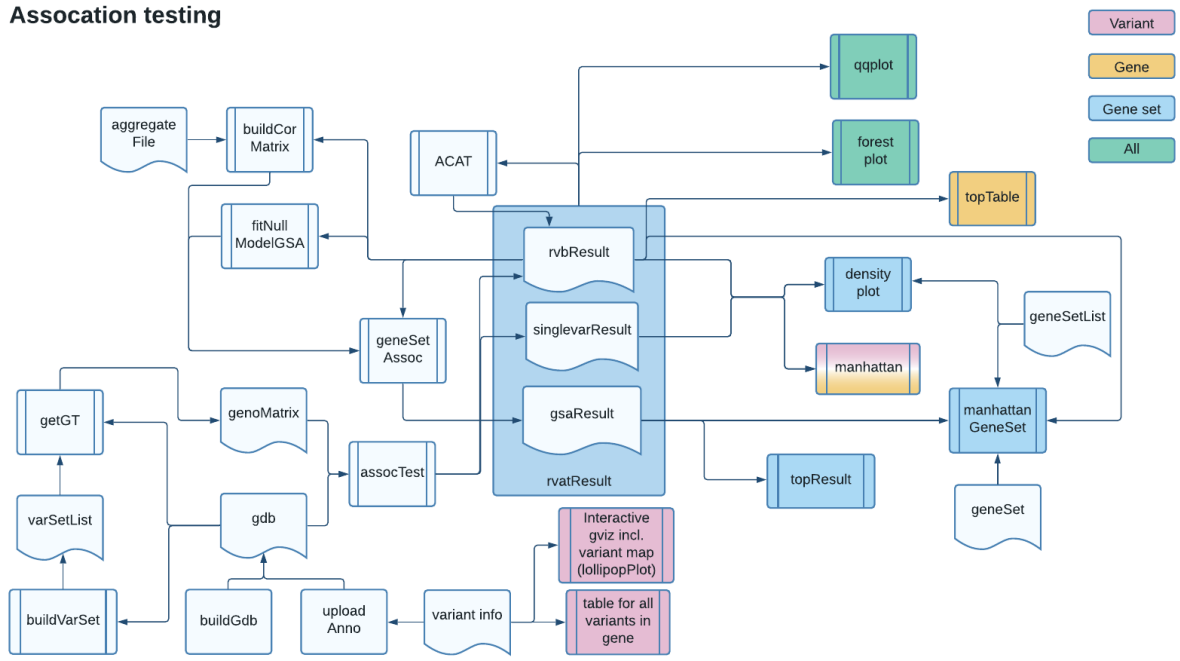
Supplementary materials

1 – Code domainVarSet()

```
domainVarSet = function(gdb,output,varSetName,unit_domainID, unitName, unitTable,
                        positions = "POS", domains, domainsID = "domain_ID",
                        domainsStart="domain_start", domainsEnd="domain_end") {

    gdb <- gdb(gdb)
    domains <- readr::read_delim(domains, col_names = TRUE)
    if (sum(c(domainsID, domainsStart, domainsEnd) %in% colnames(domains)) != 3) {
       error("The `domains` file must contain columns with domain IDs, start
             positions, and end positions. Check whether the default columns
             `domain_ID`, `domain_start`, and `domain_end` are present, or define
             similar columns that are present in the `domains` file!")
    }

    if (length(unit_domainID == 1) {
      if (substr(unit_domainID, nchar(unit_domainID)-2, nchar(unit_domainID)) %in%
          c("txt", "csv", "tsv")) {
        unit_domainID <- as.character(read.table(unit_domainID)[,1])
      }
    }

    snpEff <- getAnno(gdb, unitTable)
    if(sum(c("CHROM", positions, "VAR_id", unitName) %in% colnames(snpEff)) != 4) {
      error("The `unitTable` must contain the columns 'CHROM', `positions`,
      'VAR_id', and `unitName`. Check whether the correct `positions` and
       `unitTable` names were given!")
    }

    snpEff_gr <- GenomicRanges::GRanges(seqnames = snpEff$CHROM, ranges =
         IRanges::IRanges(start = snpEff[[positions]], width = rep(1,
             nrow(snpEff))), VAR_id = snpEff$VAR_id, unit = snpEff[[unitName]])

    snpEff_chroms <- rbind(snpEff[,c("CHROM", unitName)]) %>% distinct()
    colnames(snpEff_chroms) <- c("CHROM", "unit")

    VAR_ids_W <- unname(sapply(unit_domainID, getVARidWeights, snpEff_chroms =
                    snpEff_chroms, snpEff_gr = snpEff_gr, domains = domains, domainsID =
                    domainsID, domainsStart = domainsStart, domainsEnd = domainsEnd))

    varSetName <- rep(paste0("domains_", varSetName), length(unit_domainID))
    varSetsDF <- data.frame(unit = unit_domainID, VAR_id_W = VAR_ids_W,
                            varSetName = varSetName)
    varSetsDF <- varSetsDF[varSetsDF$VAR_id_W != "",]
    varSetsDF <- varSetsDF %>% tidyr::unite(V1, unit, VAR_id_W, varSetName, sep = "|")
    readr::write_delim(varSetsDF, output, col_names = FALSE)
  }

  getVARidWeights <- function(unit_domainID, snpEff_chroms, snpEff_gr, domains, domainsID,
                              domainsStart, domainsEnd) {
    transcript <- unlist(strsplit(unit_domainID, "_", fixed = TRUE))[1]
    if (transcript %in% snpEff_chroms$unit) {
      ranges <- GenomicRanges::GRanges(seqnames = unique(snpEff_chroms[snpEff_chroms$unit
                  == transcript,"CHROM"]), ranges = IRanges::IRanges(start =
                  domains[[domainsStart]][domains[[domainsID]] == unit_domainID,
                  domainsStart], end = domains[domainsEnd]][domains[[domainsID]] ==
                  unit_domainID, domainsEnd]))

      snpEff_gr_small <- snpEff_gr[grepl(transcript,snpEff_gr$unit, fixed = TRUE),]
```

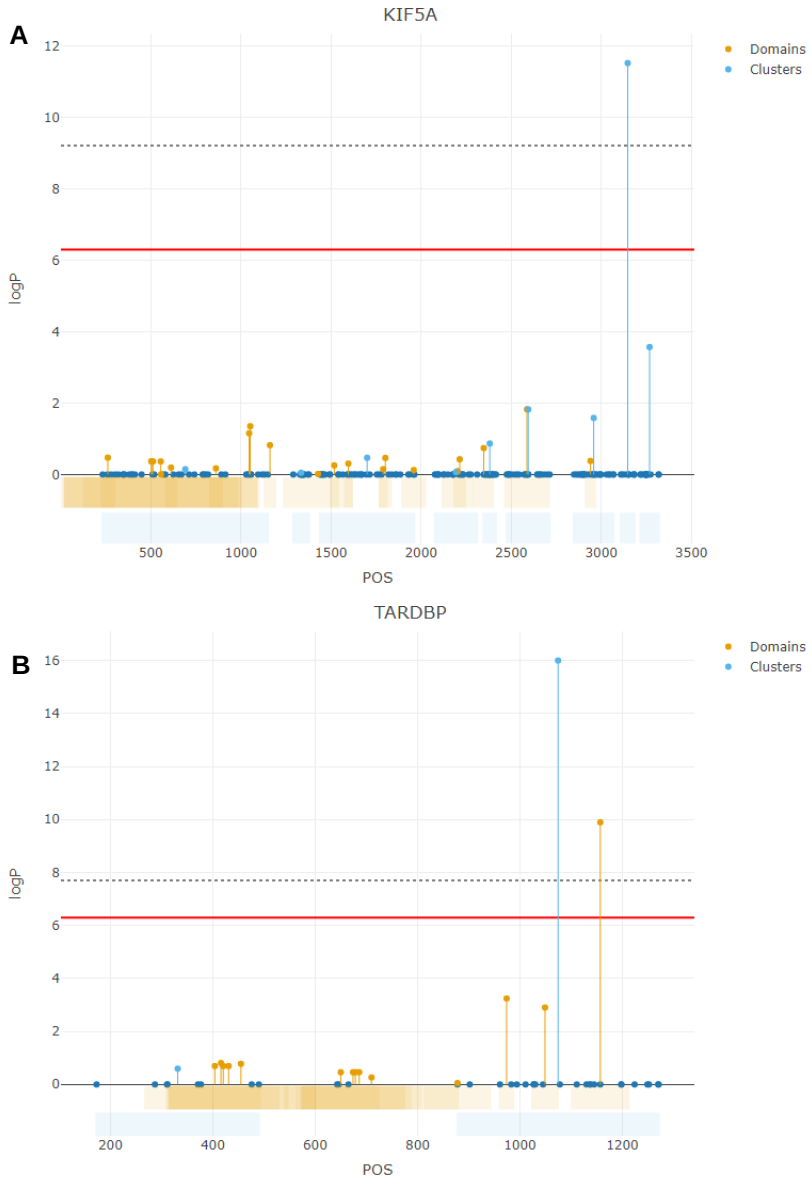
```
        sub_by_overlap <- IRanges::subsetByOverlaps(snpEff_gr_small, ranges)

        if (length(sub_by_overlap) == 0) {
          return(“”)
        } else {
          return(paste0(paste0(unique(sub_by_overlap$VAR_id), collapse = ","), "|",
                 paste0(rep(1,length(unique(sub_by_overlap$VAR_id))), collapse = ",")))
        }
      } else {return("")}
}
```

2 – Process diagram of association testing within the RVAT package



This overview shows that a *gdb* including a vcf and variant information and a *varSet* are necessary to make a *genoMatrix*, which is turn is the input of the *assocTest* function. The results of association tests and gene sets have their own class, which all fall under the class *rvatResult*. The methods for visualisation can be split into methods that were already defined before my projects and methods that we defined in this study. This last group consists of forest plots, topResult, *manhattanGeneSet*, and *densityplot*.

3 – Mutation plots *TARDBP* and *KIF5A*



A. *KIF5A* only shows evidence of a hotspot of ALS mutations for one of the spatial clusters, not in any of the domains. The P-value for the whole gene is 6.17×10^{-10} (OR = 1.31). Including only the domains results in an omnibus P-value of 0.75, while including the spatial clusters gives an omnibus P-value of 2.68×10^{-11}. Removing intermediate-frequency variants from the cluster causes an increase in P-value from 2.97×10^{-12} (OR = 1.44) to 0.054 (OR = 1.61). This implies that the signal seen in *KIF5A* is single-variant driven.

B. *TARDBP* shows strong evidence for a hotspot in both the spatial clusters and the domain-based clusters. This supports the previously reported hotspot of ALS mutations in *TARDBP*. The P-value for the whole gene is 1.995×10^{-8} (OR = 2.64), whereas the omnibus P-value of all domains and spatial clusters combined is 2.00×10^{-15}. The P-value of the cluster with all variants up to intermediate-frequency variants is 1.00×10^{-16} (OR = 9.18), and remains low with only ultra-rare variants (P = 4.53×10^{-10}, OR = 8.36). The right-most domain has a P-value of 1.28×10^{-10} (OR = 33.17) when including everything up to intermediate-frequency variants, and a P-value of 2.92×10^{-6} (OR = 20.4) when including only ultra-rare variants. This implies that the signal in *TARDBP* is driven by the whole group of variants in the cluster and domain.

4 – Variant information of spatial clusters and domains in *NEK1*

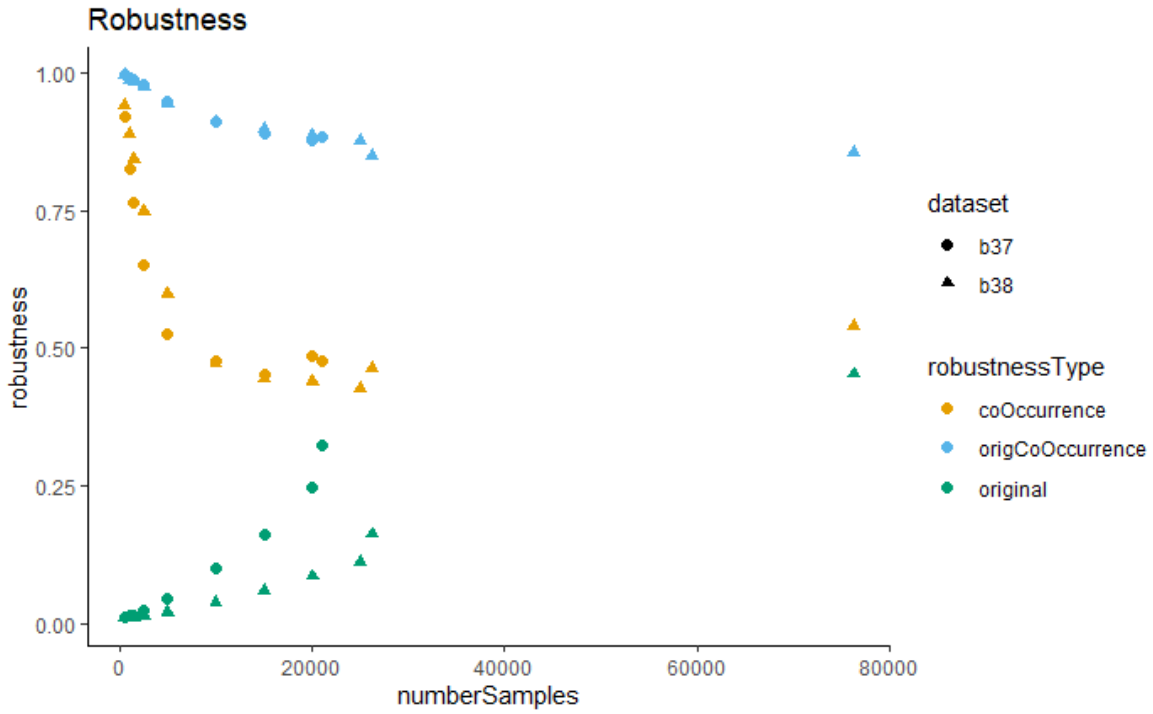
The single variant association test results were generated by Paul Hop using the statistical test 'firth'.

The cluster number correspond with the clusters seen in Figure 3. The variants in bold text are the variants driving the most significant signals in *NEK1*.

Cluster	Variant ID	MAF	Case Carriers	Ctrl Carriers	Single variant association test P-value	OR	rsID
1	6643455	4,94E-05	2	6	0,2860188	2,35336699	rs1383934287
	6643456	6,15E-06	1	0			rs1131690775
	6643457	4,28E-05	5	2	0,00619511	8,17930522	rs387906890
	6643458	6,12E-06	0	1			rs387906890
	6643459	6,09E-06	1	0			rs765039384
	6643462	6,09E-06	0	1			
	6643465	1,22E-05	0	2	0,85108613	1,35129378	rs201610555
	6643466	1,22E-05	1	1	0,32434273	3,45116133	rs756261702
	6643467	1,83E-05	1	2	0,89867716	1,148811	rs764374761
	6643470	6,19E-06	0	1			rs779421511
	6643500	3,01E-05	2	3	0,08855707	4,64268354	rs760763407
	6643501	6,03E-06	1	0			rs1313271072
	6643502	3,62E-05	2	4	0,16081711	3,3375705	rs115005766
	6643506	1,09E-04	3	15	0,85223985	0,89546006	rs201350526
	6643507	1,21E-05	0	2	0,78863138	1,55348116	rs750846188
	6643508	1,81E-05	2	1	0,05456074	7,37546962	rs781184197
	6643510	6,03E-06	0	1			rs1431339437
	6643513	1,21E-05	2	0	0,0709521	11,589821	rs780748559
	6643514	1,81E-05	1	2	0,22315799	3,95283761	rs1770058884
	6643517	1,21E-05	0	2	0,98450593	1,03071342	rs35093214
	6643519	6,05E-06	0	1			rs1282967368
	6643554	4,24E-05	4	3	0,00458326	8,35006304	rs1049502301
	6643555	1,81E-05	1	2	0,53366352	2,03273551	rs1270134755
	6643558	6,03E-06	0	1			rs1770585618
	6643559	6,03E-06	1	0			
	6643561	6,03E-06	0	1			
	6643562	6,03E-06	0	1			
	6643640	6,15E-06	1	0			rs1404362599
	6643641	6,08E-06	0	1			
	6643642	6,06E-06	1	0			rs200825809
	6643643	1,21E-05	2	0	0,0419785	13,4472848	rs200825809
	6643644	6,05E-06	0	1			rs773222357
	6643646	1,21E-05	1	1	0,31193774	3,40566545	rs749503943
	6643647	6,04E-06	1	0			rs761234040
	6643648	6,04E-06	0	1			
	6643649	6,05E-06	1	0			rs753341812
	6643650	6,09E-06	1	0			
	6643652	6,26E-06	0	1			rs1264653671
2	6643098	6,08E-06	0	1			rs925864732
	6643102	2,42E-05	1	3	0,53986975	1,8876726	rs755410424
	6643108	6,03E-06	0	1			rs1458362114
	6643112	6,05E-06	0	1			rs745711143
	6643114	6,03E-06	0	1			
	6643115	3,62E-05	4	2	0,06788285	4,63409024	rs775516158
	6643118	7,23E-05	4	8	0,19985989	2,20507727	rs10034957
	6643120	6,03E-06	0	1			rs756685207
	6643121	1,21E-05	0	2	0,93126415	1,14535724	rs756685207
	6643122	6,03E-06	1	0			
	6643123	6,03E-06	0	1			rs777255986
	6643124	1,21E-05	2	0	0,16250004	6,7584655	rs753392280
	6643126	6,03E-06	0	1			rs201587614
	6643130	2,41E-05	2	2	0,04965825	6,55662358	rs794727032
	6643131	1,21E-05	0	2	0,96526306	1,07043929	

	6643211	6,03E-06	1	0			rs1322422661
	6643213	6,03E-06	0	1			rs760983006
	6643216	**0,003690**	**174**	**438**	**5,973E-14**	**2,05984989**	**rs200161705**
	6643217	2,41E-05	3	1	0,0138901	10,3116414	rs191859401
	6643219	6,03E-06	0	1			rs142236342
	6643222	6,03E-06	0	1			rs1371101243
	6643224	6,03E-06	0	1			rs756066992
	6643225	6,03E-06	1	0			
	6643228	3,02E-05	1	4	0,4393302	2,19198808	rs765712201
	6643229	1,21E-05	2	0	0,06323551	15,2318633	rs772747361
	6643230	6,03E-06	1	0			rs772747361
	6643231	1,21E-05	2	0	0,01006538	26,2533145	rs762504963
	6643232	1,81E-04	5	25	0,82696431	1,11204831	rs61737748
	6643236	6,03E-06	0	1			rs1767386066
	6643238	6,03E-06	0	1			
	6643239	1,81E-05	0	3	0,90693246	1,19820528	rs886059235
	6643240	1,81E-05	2	1	0,01748918	12,2905109	rs1402701809
	6643245	1,81E-05	0	3	0,84671483	0,75366739	
	6643247	6,03E-06	0	1			rs1767393763
	6643251	6,05E-06	0	1			rs1451246726
	6643252	3,64E-05	0	6	0,56804405	0,4704587	rs747225801
3	6641863	1,21E-05	0	2	0,24593852	0,2113076	
	6641865	6,03E-06	1	0			
	6641867	1,21E-05	0	2	0,84422614	1,37111504	rs764463497
	6641872	1,21E-05	1	1	0,90486727	1,14885756	rs779767983
	6641874	**4,88E-04**	**42**	**39**	**2,5535E-15**	**6,39387433**	**rs199947197**

5 – Robustness of spatial clustering using variant pairs



This plot shows three trends in the robustness of a spatial clustering based on the frequency that two variants occurred in the same cluster over 500 iterations. Both the robustness in b37.75 and b38.105 were calculated at sample sizes 500, 1000, 1500, 2500, 5000, 10,000, and 20,000. For b37.75, we also took all controls and a subset of cases, which resulted in 21,014 cases. For b38.105, we also took 25,000 random samples, all cases and a subset of controls (26,256 samples) and all controls and half of the patients (76,339 samples).

The green (bottom) datapoints show the robustness of our clustering using the original calculations described by Lu et al. (2019). There is an upward trend in the robustness, which can be explained by the fact that the more variants are included, the more likely variants are to frequently co-occur in a cluster. However, overall the robustness remains low using this method because instead of altering a parameter in the clustering algorithm, we took a subset of the data. This meant that it was virtually impossible for rare variants to occur in all iterations. Subsequently, this caused the maximum number of times certain variant pairs could occur together to be lower than 500. We corrected for this in the orange (middle) datapoints by calculating the robustness of a variant pair using the maximum number of iterations in which they could occur together, instead of the total number of iterations we performed. Here, we see a downward trend towards roughly the same robustness as with the original method at higher sample sizes. This can be explained by the fact that this method calculates the robustness relative to all the clusters found in the 500 iterations, and not in relation to a reference clustering. At smaller sample numbers, variants are more likely to end up in the same cluster than they are at larger samples numbers, resulting in a higher robustness score. This is because the variants will, generally speaking, be further apart from each other, leading to a smaller gap between the mean and median distance between variants, which in turn results in fewer and thus more similar clusters over iterations. In an attempt to compare the robustness results to a reference clustering, we also calculated the robustness of only the variant pairs that were present in the original clustering with all samples included (blue/top datapoints). The robustness of these variant pairs is much higher than that of the complete list of variant pairs, implying that the clusters that are found in the cluster based on the whole dataset are more robust than the clusters that could emerge through taking a subset of the data.