BMC Medical Genomics

RESEARCH

Open Access

Identification of recurrent genetic patterns from targeted sequencing panels with advanced data science: a case-study on sporadic and genetic neurodegenerative diseases

M. Tarozzi[1], A. Bartoletti-Stella[2,3], D. Dall'Olio[4], T. Matteuzzi[4], S. Baiardi[2,3], P. Parchi[2,3], G. Castellani[2*†] and S. Capellari[3,5†]

Abstract

Background: Targeted Next Generation Sequencing is a common and powerful approach used in both clinical and research settings. However, at present, a large fraction of the acquired genetic information is not used since pathogenicity cannot be assessed for most variants. Further complicating this scenario is the increasingly frequent description of a poli/oligogenic pattern of inheritance showing the contribution of multiple variants in increasing disease risk. We present an approach in which the entire genetic information provided by target sequencing is transformed into binary data on which we performed statistical, machine learning, and network analyses to extract all valuable information from the entire genetic profile. To test this approach and unbiasedly explore the presence of recurrent genetic patterns, we studied a cohort of 112 patients affected either by genetic Creutzfeldt–Jakob (CJD) disease caused by two mutations in the *PRNP* gene (p.E200K and p.V210I) with different penetrance or by sporadic Alzheimer disease (sAD).

Results: Unsupervised methods can identify functionally relevant sources of variation in the data, like haplogroups and polymorphisms that do not follow Hardy–Weinberg equilibrium, such as the *NOTCH3* rs11670823 (c.3837 + 21 T > A). Supervised classifiers can recognize clinical phenotypes with high accuracy based on the mutational profile of patients. In addition, we found a similar alteration of allele frequencies compared the European population in sporadic patients and in V210I-CJD, a poorly penetrant *PRNP* mutation, and sAD, suggesting shared oligogenic patterns in different types of dementia. Pathway enrichment and protein–protein interaction network revealed different altered pathways between the two *PRNP* mutations.

Conclusions: We propose this workflow as a possible approach to gain deeper insights into the genetic information derived from target sequencing, to identify recurrent genetic patterns and improve the understanding of complex

*Correspondence: gastone.castellani@unibo.it
†G. Castellani and S. Capellari have equally contributed to the work
[2] Department of Experimental, Diagnostic and Specialty Medicine, University of Bologna, Bologna, Italy
Full list of author information is available at the end of the article

© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

diseases. This work could also represent a possible starting point of a predictive tool for personalized medicine and advanced diagnostic applications.

Keywords: NGS, Genetic modifiers, Polygenic score, Gene panels, Machine learning, Complex diseases, Neurodegeneration, CJD, Alzheimer's Disease

Background

Gene panels are a powerful clinical and research tool that allow to perform massively parallel sequencing on a set of genes of interest. This technology is often used in the clinic practice as a diagnostic tool, however, at present a large fraction of the collected genetic information remains unexploited, since their valence is difficult to assess. On the research side however, genetic modifiers and oligogenic patterns of inheritance are gaining an increasing interest, because of the phenotypic variability of some diseases and as a possible answer to missing heritability of other conditions [1]. An increasing number of papers is pointing out that the genetic part of the missing information about heritability and phenotypic heterogeneity is likely to be addressed by a set of variants reinforcing themselves in connected molecular pathways rather than a specific mutation yet to be discovered [1–3]. It is therefore of great relevance to improve methods of analysis to acquire a richer insight of the overall genetic features. For this reason, in the genomic field there is an increasing use of machine learning methods (ML) and network analysis, which allow to identify recurrent genetic patterns in the data and to integrate and amplify single genetic variants in their biological context [4–10]. Neurodegenerative brain diseases are progressive and fatal conditions primarily affecting the central nervous system. In the last decades, linkage studies in families with a disease showing Mendelian inheritance identified high-penetrant mutations in causal genes in a minority of them. In the vast majority, common variants in genes with significant associations in genome-wide association studies (GWAS) concurred, with a modest increase in disease risk, to the disease. Genetic risk factors or modifiers play an important role both as additional risk factors in co-occurrence with incompletely penetrant mutations but also as modulators of disease severity, age of onset and in the overall course of the disease [11–15]. Here, we consider three inheritance models in two neurodegenerative diseases, genetic Creutzfeldt–Jakob Disease (gCJD) and sporadic Alzheimer Disease (sAD). Creutzfeldt–Jakob Disease (CJD) is the most common human prion disease [16], where genetic forms caused by mutations in the *PRNP* gene account for 15% of the cases and show autosomal dominant inheritance with variable penetrance [17]. We focused on the two most common *PRNP* mutations in the Italian population, the highly penetrant p.Glu200Lys (E200K group) and the p.Val210Ile (V210I group), that shows low penetrance [18]. The *PRNP* gene, located in chromosome 20 in the human genome, is 16 Kb long and made up of two exons, the second containing the whole open reading frame, resulting in a mature protein of 208 amino acids. The most important known risk factor and phenotypic modifier is the polymorphism at the codon 129 of the *PRNP* gene, that can result either in Methionine or Valine. Homozygotes are overrepresented in the population affected by prion diseases, while heterozygosity has a protective role. Sporadic Alzheimer Disease is the third model considered in this work: sAD is known to be influenced by both genetic and environmental factors. Extensive studies led to the discovery of important predisposing factors, such as *APOE* genotype, and variants in *ABCA7, SORL1, TREM2* genes [19]; nevertheless the missing heritability of sAD remains an important open question [20, 21]. In this study we tried to improve current approaches towards target sequencing data analysis considering each single nucleotide variant (SNVs) and small indels obtained through DNA target sequencing of twenty-nine genes known to play a role as risk factors or determinants of dementias, on one-hundred and twelve patients affected by either sAD or gCJD caused by either the highly penetrant mutation p.Glu200Lys or a lowly penetrant p.Val210Ile mutation. This study employs a data analysis workflow involving a combination of statistical, machine learning and network analysis to extract all the valuable information to identify and evaluate potential polygenic contributions to neurodegenerative dementia. We focused on differences of recurrent genetic patterns covered by our gene panel between groups of interest, sAD vs gCJD, and in the CJD group between the two described mutations, p.Glu200Lys and p.Val210Ile. As the results of our case study show, this workflow represents a suitable approach to acquire a deeper understanding of the genetic pattern present in target sequencing data, able to improve our understanding of the underlying molecular biology of complex diseases and a possible starting point as a predictive tool for personalized medicine applications.

Materials and methods

Subjects

We recruited patients with definite, probable, probable laboratory-supported, and possible CJD or AD diagnosed

according to National Institute of Aging/Alzheimer's Association (NIA/AA) [22] or International Working Group-2 (IWG-2) [23] for AD and updated clinical diagnostic criteria for sporadic Creutzfeldt–Jakob disease [24], afferent to the Cognitive Disorders and Dementia Center of the UOC Clinica Neurologica, Bologna, either as outpatients, inpatients, or sent for genetic analysis between 2010 and 2019. One-hundred-twelve patients with either gCJD (n = 66) or sAD (n = 46) were recruited. Among the sixty-six gCJD patients, forty were carriers of the p.Val210Ile and twenty-six of the p.Glu200Lys mutation. For brevity, in this work we will refer to these groups of patients as V210I and E200K groups, whilst specific protein coding variants will be reported with the aminoacidic shift nomenclature and non-coding variants with nucleotide shift nomenclature. Ethical approval was obtained from the ethical board of our institution. For all subjects, written informed consent was provided. All methods were performed in accordance with the relevant guidelines and regulations.

DNA extraction

Genomic DNA from peripheral blood was extracted using the Maxwell 16 extractor (Promega, Madison, WI, USA) and quantified using the Quantus Fluorometer (Promega) with QuantiFluor double-stranded DNA system.

Target sequencing and secondary analysis

Target sequencing covers 29 genes (Additional file 1: Table S1 see also Bartoletti-Stella et al. 2018 [25]), known to play a role as risk factors or primary determinants in different types of dementia [26]. Libraries were constructed with the amplicon-based assay TruSeq Custom Amplicon v1.5 (TSCA, Illumina, CA, USA), sequencing was performed on a MiSeq sequencer using Illumina V2 reagent kit, using 2×150 bp paired end read cycles. Raw data were analyzed by the MiSeq Reporter software (Illumina), aligned to GRCh37/Hg19 using bwa-mem with variant calling and depth of coverage calculation with Genome Analysis Toolkit (GATK) [27]. During the variant calling steps, variants were filtered based on quality using Q30 as threshold, which means that at most 0.1% error rate is allowed.

Data transformation

To obtain an input suitable for the computational and statistical analysis, containing the whole genetic variability in the dataset and still maintaining the single-patient detail, the genetic information contained in the Variant Call Format (VCF) files was transformed into binary data through an in-house Python script. Our script generates a matrix in which each row represents a variant reported in the provided VCF files at least once and each column is named after an ID assigned to each patient. In the matrix, 0 indicates that the variant is not present in the VCF file of the patient whereas 1 indicates its presence. A second version of the script was also used to produce a ternary matrix in which the zygosity information was added, thus 1 indicates heterozygosity and 2 homozigosity. On these matrices, machine learning methods and statistical analysis were applied using scikit-learn [28], seaborn [29] and plotly express [30] packages on Jupyter notebooks.

Machine learning analysis

We used both supervised and unsupervised methods to extract as much valuable information as possible from our transformed data. To visualize such high dimensional data, we tested different dimensionality reduction techniques, such as Principal component analysis (PCA), t-distributed stochastic neighbour embedding (t-SNE) using Jaccard similarity as metric. As supervised methods, we used decision trees on binary data labelled accordingly to the disease or genotype (sporadic, p.Glu200Lys, p.Val210Ile) of each patient. The classifier was trained on a random selection of 2/3 of the dataset and adequate branching depth was set to avoid overfitting. The classification rules were tested on a validation set represented by the remaining 1/3 of the dataset.

Statistical analysis

Allele frequencies of each variant found at least in one patient of our cohort were obtained by the ternary matrix in which the zygosity information was included. We then compared allele frequencies calculated in the sAD and in the gCJD populations with those reported in the gnomAD database [31] for the non-Finnish European population using Fisher's exact test and Benjamini–Hochberg multiple test correction. We defined *MAPT* haplotypes in our population using the two coding SNV rs1052553 and rs1800547 [32, 33], and we tested for Hardy–Weinberg equilibrium using the R package "HardyWeinberg" [34].

Protein interaction network and pathway analysis

To further explore the biological interplay of genes harboring significant variants we performed protein–protein interaction network and pathway enrichment analysis. Protein–protein interaction network (PPIn) was built by merging data from four state-of-the-art protein–protein interaction databases [35–38]. All of them were obtained with high-throughput assays and comprise only biophysical interactions (i.e., molecular docking) between proteins [39]. The network resulting from their union covers almost 14,000 genes and 110,000 interactions between gene products. For each disease group, we built a

group-specific subnetwork by mapping on the PPIn those genes harboring at least one variant with a p-adj < 0.05 and considering their nearest neighbor genes. We then focused on differences between group-specific subnetworks by identifying, for the group pairs of interest (i.e., sAD-gCJD, V210I-E200K), the sets of genes unique to one disease with respect to the other. On this sets of genes, we performed pathway enrichment using Gene Ontology [40, 41], to explore which pathways are likely to be affected by differences in genes harboring at least one statistically significant (p < 0.05) variant between disease groups.

Results

Unsupervised methods identify functionally relevant genetic modules

Binary transformation of the genetic information contained in VCF files led to the generation of a matrix with shape 1046X112 where each row identifies a variant, and each column identifies a patient. The first exploratory analysis of the 1046X112 matrix containing the whole genetic information was performed trough dimension reduction with Principal Component Analysis (PCA) (Fig. 1). In the 2D plot each dot represents a patient. The first two principal components (PC1 and PC2) explain 22% of the overall variance of the dataset (Additional file 1: Table S3). PC1's main contributors are a group of SNPs that are all harboured in the *MAPT* genomic region (Additional file 1: Table S2), that previous works have defined as haplotype-specific [33, 42]. The second principal component involves a more heterogeneous group of SNPs in which the SNP rs11670823 in the *NOTCH3* gene is the major contributor (Additional file 1: Table S3).

We labelled each sample according to the disease affecting the patient, the genotype (sporadic, p.Glu200Lys, p.Val210Ile) and a label marking a possible batch effect due to different sequencing runs. None of these labels matched the identified clusters (Additional file 1: Fig. S1). Based on the loadings and score values of the PCA we focused on the main genetic sources of variation in the dataset. We identified the two main *MAPT* haplotypes (H1,H2 and H1/H2), according to two coding SNPs rs1052553 and rs1800547 [33, 42] which are in linkage disequilibrium (LD) with the rs11575896 (first contributor to PC1, Additional file 1: Table S2). The result of the labelling of our dataset according to *MAPT* haplotypes perfectly matches the clusters in the PCA plot, as shown in Fig. 1. Our population is in Hardy–Weinberg (HW) equilibrium for the tested SNPs. The distribution on the y-axis recognizes specific SNPs patterns associated to haplotypes of *NOTCH3* (Additional file 1: Fig. S2). [43] Interestingly, the SNP rs11670823 (c.3837 + 21 T > A), which is the major contributor to the PC2, is in LD with three *NOTCH3* haplotype defining SNPs (rs1044009, rs104423702 and rs4809030) [43], and is in HW disequilibrium (p = 0.03) in the complete cohort (sAD p = 0.085, E200K p = 0.276, V210I p = 0.420).

Supervised methods recognize clinical phenotypes with high accuracy

Supervised classifiers were used for automatic recognition of genetic patterns among the 1046 variants identified in this dataset. In the 112X1046 matrix, to each sample a label corresponding to the disease (class: "CJD" or "AD") was added. The classification was achieved perfectly, with 100% of accuracy (ratio of correctly predicted observation to the total observations) on the test set, basing the classification on the two disease-causing mutations p.Val210Ile and p.Glu200Lys (Fig. 2).

To test for the presence of additional recurrent genetic patterns that could characterize a homogeneously phenotypic group and possibly act as modifier, we removed from the input data provided to the classifier only the two rows of the 1046-rows matrix indicating the disease-causing mutations. As expected, accuracy decreased both in training set and in test set, but interestingly the classifier managed to distinguish the two diseases with a good accuracy (training = 0.97, test 0.78) (Table 1).

The classification is based on eight variants involving six different genes (Fig. 3). All considered variants were reported in common databases and genomic search engines such as VarSome [44], OMIM [45], ClinVar [46] or HGMD [47] and their consequence was assessed as known disease-causing variant, risk factor, variant of uncertain significance (VUS) or benign according to the ACMG guidelines for interpretation of sequence variants [48]. Five variants are predicted to be benign and are intronic or synonymous, three of them are classified as variants of uncertain significance and are missense or located in 3'UTR regions.

Statistical analysis of variants frequency

For each of the 1046 variants detected, allele frequency was calculated. We calculated separately allele frequencies in the sAD and in the gCJD group. The latter was further divided according to the presence of the p.Glu200Lys or p.Val210Ile mutations. Each allele frequency was then compared to those reported into the GnomAd database [31] for the European (non-Finnish) population. Differences between observed and expected allele frequency were tested for statistical significance with Fisher's Exact test and Benjamini–Hochberg multiple test correction. Table 2 summarizes the number of each type of variants per group and show the average number of variants per patient in the different classes (gene list reported in Additional file 1: Table S4).

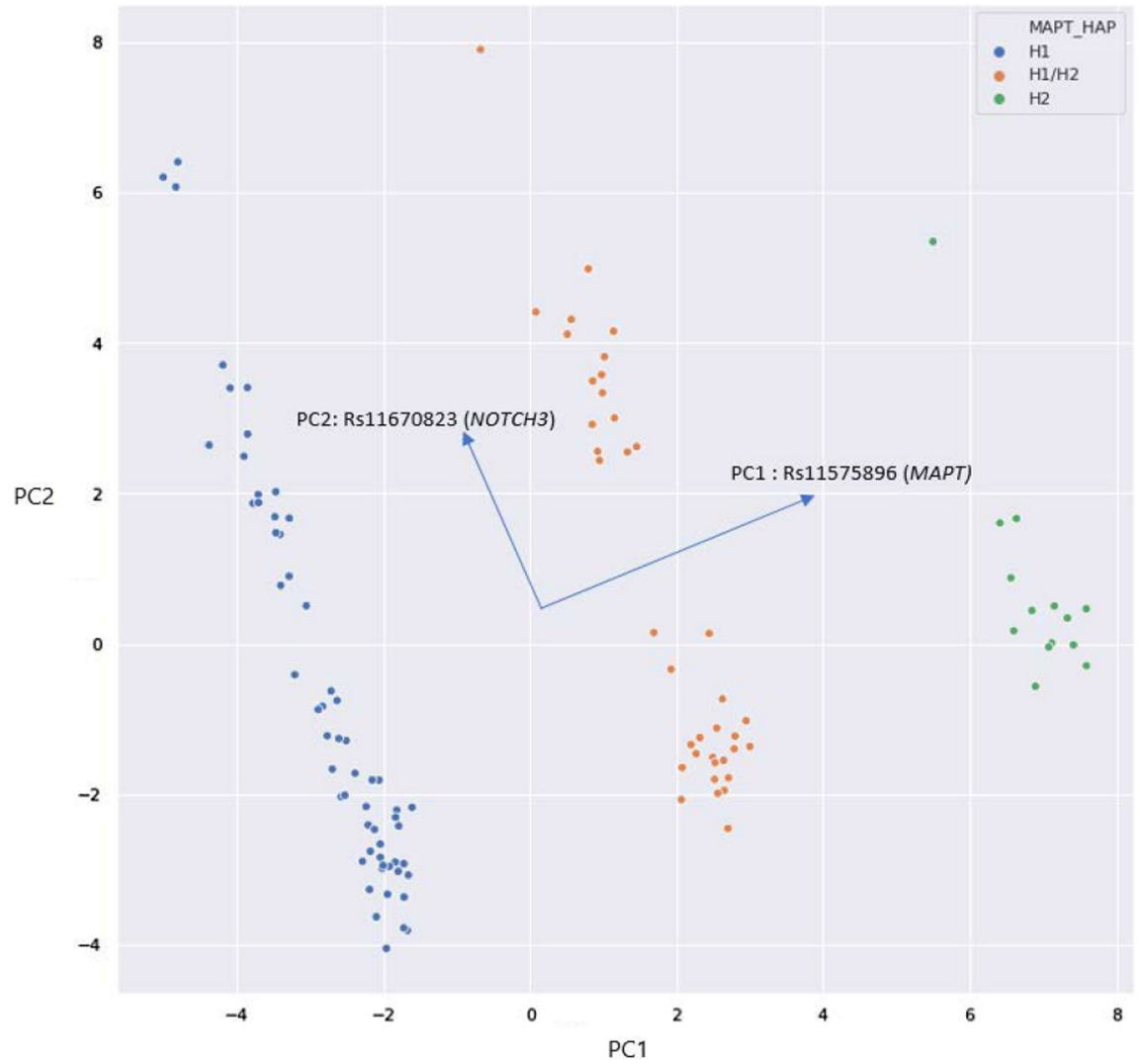

Fig. 1 2D plot of the Principal Component Analysis (PCA) computed on the 1046 × 112 ternary matrix. PCA is a dimensionality reduction technique that computes an orthogonal linear transformation of the data to a new 2D coordinate system so that the greatest variance is on the x-axis (PC1) and the second greatest variance on y-axis. Each dot represents a patient, that is plotted in the 2D space accordingly to its genetic profile expressed in the ternary matrix. PC1 and PC2 show the main sources of variance in our data, accounting for 22% of overall variance, that are represented by variants on *MAPT* and *NOTCH3* genes, respectively. PCA plot and hierarchical clustering recognize clusters that correspond to the *MAPT* haplotypes on the x-axis, as shown by coloured labels in the picture legend. Similarly, the distribution along the y-axis matches haplotypes in the *notch3* gene (not shown)

Pathway analysis and protein–protein interaction network

To have functional insights of the consequences of the alterations in allele frequencies, genes harbouring at least one variant with $p < 0.05$ were used as input for pathway analysis with GO database (Fig. 4) and protein–protein interaction (PPI) network. Since part of the affected pathways are shared among the considered conditions, results are reported as differences between comparisons of two groups. Comparison of sAD vs gCJD in the PPI network shows a clear centrality of interactions of *APP*, *PSEN2* and *APOA1* in the AD but not in the CJD group (PPIn tables and figure in Additional file). Functional analysis of the same coupled comparison points out a significant ($p < 0.05$) enrichment in the sAD group (compared to gCJD) of the GO terms involving regulation of the apoptotic signaling pathway, supramolecular fiber organization, antigen processing and presentation of exogenous peptide antigen. Interestingly, in the CJD group we found and enrichment of GO terms involving the ER responses to stress, protein folding, regulation of

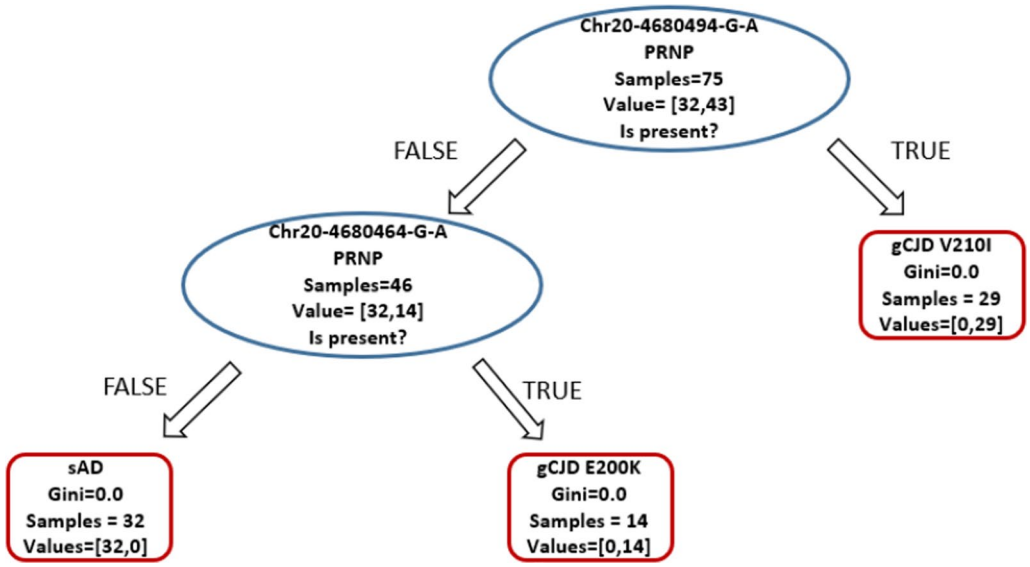

Fig. 2 Dataset classification according to decision trees analysis: this supervised method computes on the 1046 × 112 matrix a classification based on the labels provided. The classifier correctly identifies the two disease groups on the two disease-causing mutations

Table 1 Classification metrics.

	Precision	Recall	F1	Support
sAD	0.71	0.8	0.71	15
gCID	0.85	0.77	0.85	22

Precision is the ratio of correctly predicted observation to the total predicted positive observations (TruePositive/TruePositive + FalsePositive), Recall is the ratio of correctly predicted positive observations to all observations in actual class (TruePositive/TruePositive + FalseNegative), F1 Score is the harmonic mean of Precision and Recall (F1 Score = 2*(Recall * Precision) / (Recall + Precision)). Support indicates class numerosity

mRNA maturation and splicing and in the regulation of catabolic processes. We then investigated whether functional differences within the gCJD group could provide further understanding of the different penetrance of the two mutations. In the coupled comparison between V210I and E200K, we found that only in the V210I group there is an enrichment of GO terms referring to proteasome mediated catabolic processes and antigen processing and presentation. In the PPIn results for the comparison V210IvsE200K, *APOA1* and *MAPT* together with *DCTN1* represented hubs of the network, highlighting a similarity between the enriched modules in the networks of the lowly penetrant p.Val210Ile mutation and the one of sAD, with numerous interactions and shared nearest neighbours involved in the enriched pathways. In the E200K group compared to the V210I we found a significantly altered regulation of mRNA and splicing, reflected in the PPI network by the abundant presence of members of the family of heterogeneous nuclear ribonucleoproteins (*hnRNPs* gene family) as nearest-neighbours of the input genes, in addition to the alteration of actin filament organization.

Discussion

In this work, we addressed the challenge of exploring the complete genetic information carried by target sequencing data to acquire deeper insights in the genetic contributors of complex diseases. For this purpose, we selected a population of one-hundred and twelve patients affected by two neurodegenerative diseases: sporadic (sAD) and genetic Creutzfeldt–Jakob Disease (gCJD), either due to a highly (p.Glu200Lys) or a lowly (p.Val210Ile) penetrant mutation. As supported by previous research about genetic modifiers in this field, it is possible that other factors reinforce its pathogenic role in carriers of the *PRNP* p.Val210Ile who indeed develop CJD. In sAD, no specific causative mutations are present, nevertheless GWAS have revealed many loci of common genetic variation that confer risk for developing the disease and evidence supports a polygenic contribution to disease risk from common genetic variants [13, 49–51]. Our approach is based on a binary transformation of each detected variant. On the resulting matrix we applied statistical, supervised and unsupervised machine learning methods and network analysis as an unbiased approach to discover recurrent patterns and possible genetic modifiers among the genes included in the target sequencing. In this work we refer to patients' groups as V210I, E200K and sAD groups, whilst specific protein coding variants are reported with the aminoacidic shift nomenclature and non-coding variants with nucleotide shift nomenclature.

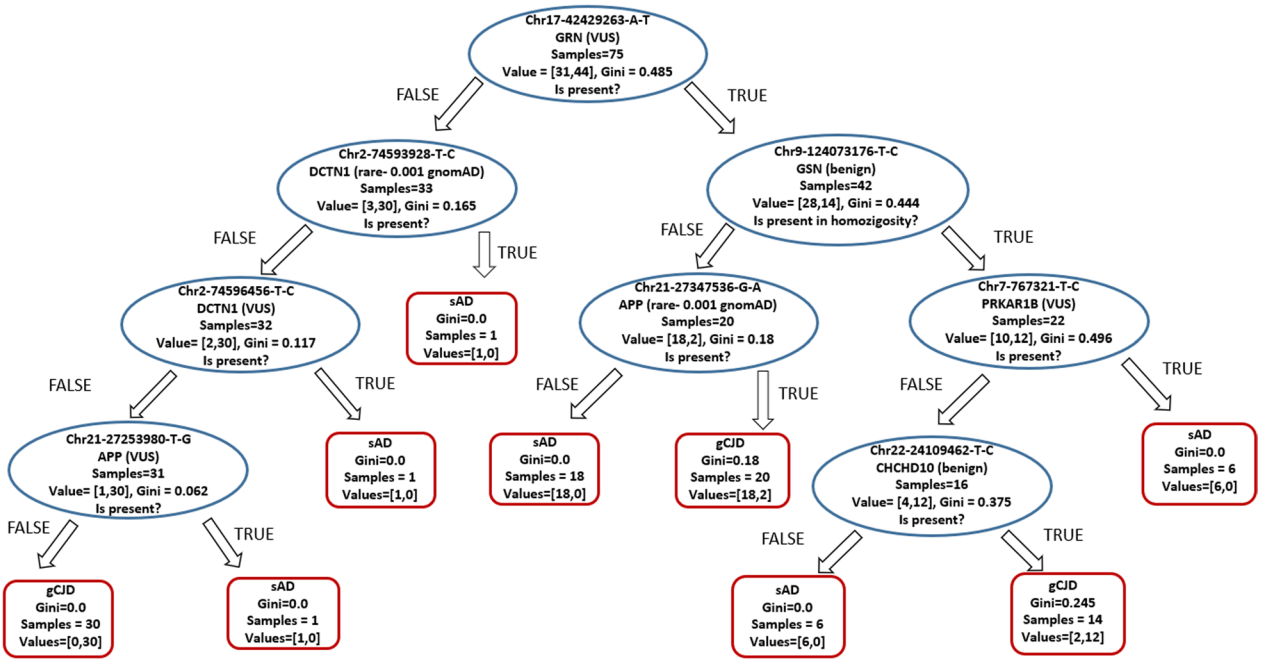

Fig. 3 Result of Decision Trees analysis on the dataset deprived of the information about gCJD-causing mutations. Classification is accomplished with 0.71 accuracy for sAD and 0.85 for gCJD. Classification is based on the reported eight variants harboured in six genes. Four of these are variants of uncertain significance not reported in the GnomAD database harbored in the genes *APP* c.*1A > C (rs748508166), GRN c.1179 + 100A > T, *DCTN1* p.Lys519Glu, *PRKAR1B* c.595 + 369 T > C (rs1342588350), two of them are rare (Minor Allele Frequency < 0.05) variants in the European population, *APP* p.Phe435 = (rs148180403, MAF = 0.001), *DCTN1* p.Ala816 = (rs1130484, MAF = 0.007) and two are common benign variants in *CHCHD10* (c.261 + 99A > G) and *GSN* (c.666 + 53 T > C). "Value" indicates the number of samples at the given node that fall into each category. The "Gini" score quantifies the purity of the node/leaf, when greater than zero implies that samples contained within that node belong to different classes while a gini score of zero means that within that node only a single class of samples exist

Table 2 Summary of results of statistical analysis on each variant detected in our target sequencing panel

Disease group	Average number of SNV per patient	Unique SNV per disease group	Unique non-synonimous SNV per disease group	Unique SNV p < 0,05 per disease group
AD (46)	145.05	654	27	72
CJD (66)	134.87	768	11	33
E200K (26)	138.73	483	14	52
V210I (40)	135.73	645	27	75

Rows identify pathologic groups with their numerosity reported between brackets. The first column shows the average number of variants carried per patient in the different disease groups. The second column shows the overall number of different variants detected in each group in at least one patient. The third column indicates variants annotated as missense, splice variants or 3'or 5' UTR in each disease group. The last column contains the number of variants with a p < 0.05 after Fisher's exact test and Benjamini–Hochberg correction despite of their annotation

(See figure on next page.)

Fig. 4 Result of functional enrichment analysis performed on genes harbouring variants with significantly altered allele frequency compared to European population reported in the GnomAd database. Results of pathway analysis are reported as significantly (p < 0.05) enriched pathways in the first group but not in the second of each coupled comparison. Since part of the affected pathways are shared among the considered conditions, results are reported as differences between comparisons of two groups. Complete results of the functional analysis with Gene Ontology and of the Protein–Protein Interaction networks are reported in Supplementary materials

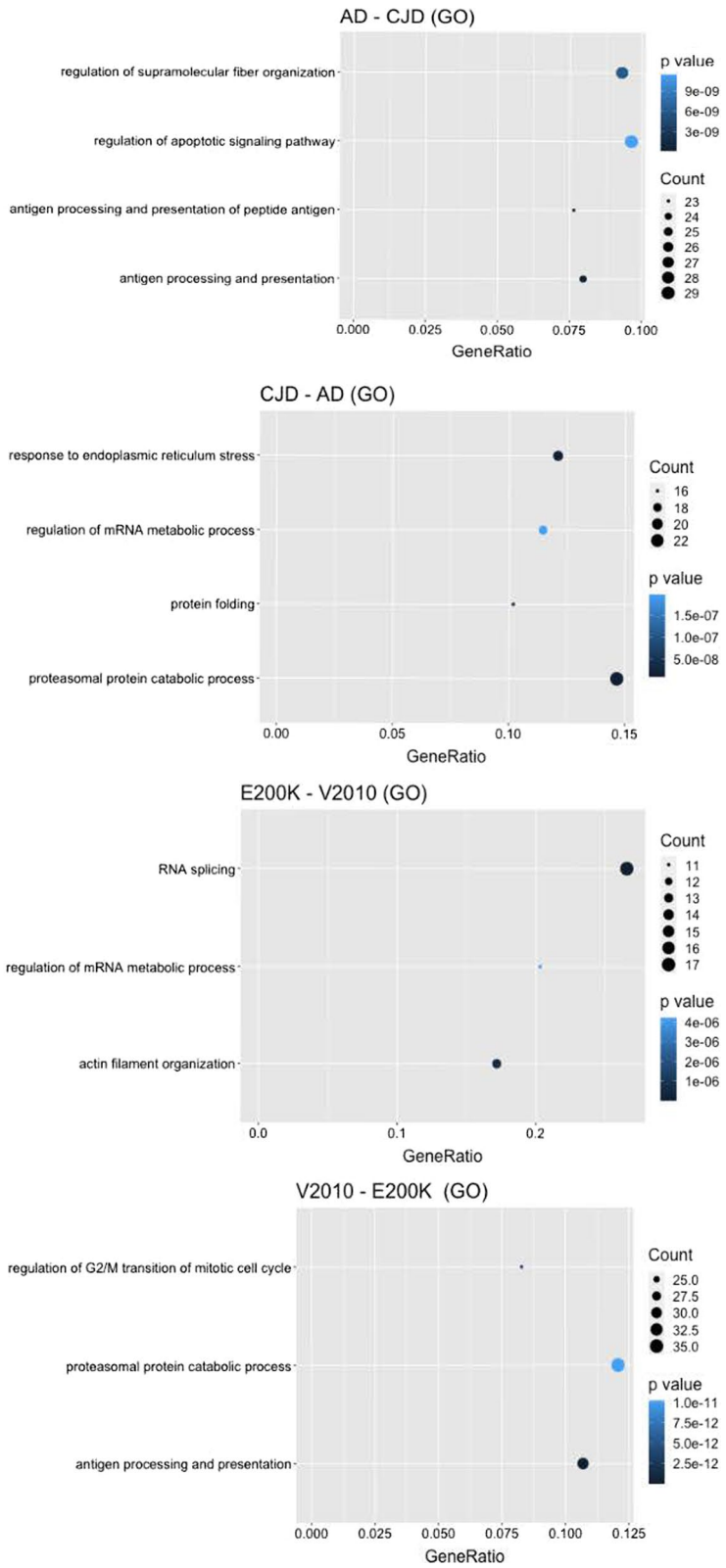

Fig. 4 (See legend on previous page.)

Unsupervised machine learning methods

Unsupervised methods identify as the main sources of variation in the dataset the haplotypes in *MAPT* and *NOTCH3* genes. Specifically, a set of haplotype-defining SNPs in *NOTCH3* do not follow the Hardy–Weinberg equilibrium law in our complete cohort suggesting a role of *NOTCH3* in the analysed neurodegenerative diseases. This role seems to be more stressed in the sAD group, both in the increasing number of variants with altered allele frequency (22 SNPs) and in the p-values when tested for HW equilibrium ($p = 0.085$). This result is in line with the functional role of *MAPT* haplotypes in most neurodegenerative diseases [43, 52–57] and with the role that *NOTCH3* in AD has been addressed by multiple previous reports [53, 58–61]. Thus, these results support the validity of our approach. Nevertheless, it must be considered that these results are dependent on the selected genes of the target sequencing and on the amount of SNVs present in those genes.

Supervised machine learning methods

Supervised methods correctly classified our samples according to the two causative mutations responsible for the genetic forms of CJD. When applied to the data deprived of the two causative mutations, decision trees classified the phenotypic groups with 78% accuracy according to eight variants (Fig. 3 and Table 1). To our knowledge, none of the variants have been previously linked to the considered conditions. Four of these variants were not previously reported in the GnomAD database in the genes *APP,* c.*1A > C (rs748508166), *GRN,* c.1179 + 100A > T, *DCTN1,* p.Lys519Glu, *PRKAR1B,* c.595 + 369 T > C (rs1342588350), two of them are rare variants in the European population, *APP* p.Phe435 = (rs148180403, MAF = 0.001), *DCTN1* p.Ala816 = (rs1130484, MAF = 0.007) and two are common benign variants in *CHCHD10* (c.261 + 99A > G) and *GSN* (c.666 + 53 T > C). Six variants were carried only by patients with either sAD or gCJD (gini = 0.0): of these, five rare or VUS variants were found only in a subset of sAD patients (the two variants in *APP,* the two in *DCTN1* and the one in *PRKAR1B*), while the benign variant in *CHCHD10* was found only in gCJD. Despite the lack of statistical power of the study, it is reasonable that at least some of these variants could play a role as a contributor to the disease risk, given that both CJD mutations are not completely penetrant [62] and the polygenic nature of sAD [21]. Decision trees have been recently proposed as a suitable method in clinical applications and precision medicine for interpreting the role of genetic variants in complex diseases [63]. Our results reinforce the importance of this supervised method to improve understanding of the role of the numerous variants of uncertain significance and as a promising path towards precision medicine applications. In addition, our results indicate that decision trees can provide accurate classification on high-dimensional genomic data.

Statistical and functional analysis

Statistical analysis performed on detected allele frequencies compared to those reported in the gnomAD database for the European non-Finnish population identified 33 to 75 variants with significantly altered allele frequency in each studied group. Each group showed a unique set of significant variants in the tested genes. Coherently with the hypothesis of a polygenic contribution in sAD, we found a higher number in sAD patients compared to gCJD both in the average number of variants per patient, with on average 145 variants carried by patients with sAD compared to the 134.87 in the genetic CJD group, and in the overall number of SNV with a significantly altered allele frequency (72 in sAD, 33 in gCJD, see Table 2 and Additional file 1: Table S4). These genes were used as input to perform functional analysis with Gene Ontology (Fig. 4) and PPI network comparing groups of interest, namely the two disease groups gCJD-sAD and the V210I-E200K CJD. In the first comparison, the PPI network identified important hubs only in the sAD group in correspondence of crucial genes in AD such as *APP*, *PSEN2* and *APOA1* despite the AD cohort did not bear any causative mutation. These results, together with the GO terms "regulation of the apoptotic signaling pathway", "supramolecular fiber organization", "antigen processing" and "presentation of exogenous peptide antigen" enriched in the functional analysis, are in line with previous reports about the polygenic nature of sAD and with its impaired pathways [50, 64–66]. With the same approach, in the CJD group we found an enrichment of pathways reported in previous functional studies as altered in this pathology, such as endoplasmic reticulum impairment, protein folding and regulation of mRNA maturation [67–70]. These results prove the validity of this approach to handle and valorise the great amount of information contained in target sequencing data and to acquire new insights about new putative risk variants. Within the CJD group, we observed differences in the lists of genes carrying altered allele frequencies, that were reflected in the functional analysis. The V210I-E200K coupled comparison showed differences between the genetic background of the same pathology triggered by different mutations. In the E200K-CJD compared to the V210I-CJD we found an altered regulation of mRNA and splicing, reflected in the PPI network by the abundant presence of members of the family of heterogeneous

nuclear ribonucleoproteins (*hnRNPs* gene family) as nearest-neighbours of the input genes, in addition to the alteration of actin filament organization. Interestingly, a similarity between the affected pathways in V210I-CJD and sAD emerged: these two groups show a high number of variants with altered allele frequencies, 75 and 72, that lead in both cases in significant alterations in pathways involving proteasome-mediated catabolic processes, antigen processing and presentation, and PPI networks sharing various hubs, such as *APOA1* and *DCTN1*. These results are in line with several previous works that claim a complex genetic background in which the reinforcing role of several variants acting together increases the risk of developing a disease both in sporadic and in genetic forms [21, 59, 67, 71]. Here, functional pathway enrichment analysis and protein–protein interaction network showed a significant alteration in genes involved in immunity, catabolic processes, RNA splicing and cytoskeletal structure maintenance. These pathways are known to be altered in both AD and CJD [66, 68, 72–74] and in other neurodegenerative conditions, suggesting a contribution of those variants in exacerbating the pathologic alteration in those pathways.

Conclusions

This work proposes an innovative approach towards the analysis of targeted NGS data, based on a binary transformation of the detected variants, on which an unbiased analysis is performed through statistical, machine learning and network analysis. Our results show that this method is a valuable workflow to explore recurrent genetic patters in homogenous phenotypic groups and increase our understanding of complex diseases. This approach can also be used to acquire new hints to identify specific SNV that could act as modifiers or risk factors in the studied condition. Specifically, we showed in our cohort how unsupervised methods can identify functionally relevant sources of variation in the data and that supervised classifiers can recognize clinical phenotypes with high accuracy based on the mutational profile of patients, thus representing a possible starting point for advanced diagnostic tools. Statistical, functional and network analysis provided functional insights that showed reliability in identifying both important known molecular features of the considered diseases and providing new insights on putative new genetic contributors. To conclude, we propose this workflow for an advanced analysis of target sequencing data in complex diseases.

Abbreviations

NGS: Next generation sequencing; ML: Machine learning; SNV: Single nucleotide variant; SNP: Single nucleotide polymorphism; SAD: Sporadic Alzheimer Disease; GCJD: Genetic Creutzfeldt–Jakob disease; MAF: Minor allele frequency; PPIn: Protein–protein interaction network; PCA: Principal component analysis.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12920-022-01173-4.

Additional file 1. Supplementary material and results: Identification of recurrent genetic patterns from targeted sequencing panels with advanced data science: a case-study on sporadic and genetic neurodegenerative diseases.

Acknowledgements

The authors are grateful to the patients and their families.

Authors' contributions

MT data analysis, drafting and revising the manuscript; ABS sequencing experiments; DD and TM contributed to data analysis; SB, PP and SC data acquisition, GC, SC and MT design of the work and revision of the manuscript. All authors contributed to the article and approved the final manuscript. All authors read and approved the final manuscript.

Funding

This work is funded by the University of Bologna, the IRCCS Institute of Neurological sciences of Bologna, and by the European Grants H2020 GenoMed4All [AM1] (Grant N. 101017549) and H2020 MSCA-ITN IMforFUTURE (Grant N. 721815).

Availability of data and materials

The VCF files supporting the conclusions of this article are available in the European Variation Archive-EMBL-EBI (https://www.ebi.ac.uk/eva/), Project: PRJEB47822, Analyses: ERZ3614541 (https://www.ebi.ac.uk/ena/data/view/PRJEB47822).

Declarations

Ethics approval and consent to participate

Ethical approval was obtained from the ethical board of our institution (Institute of neurological sciences of Bologna (IRCCS)). For all subjects, written informed consent was provided.

Consent for publication

Not applicable.

Competing interests

The authors have no competing interests to declare.

Author details

[1]Department of Medical and Surgical Sciences, University of Bologna, Bologna, Italy. [2]Department of Experimental, Diagnostic and Specialty Medicine, University of Bologna, Bologna, Italy. [3]IRCCS Institute of Neurological Sciences of Bologna, Bologna, Italy. [4]Department of Physics and Astronomy, University of Bologna, Bologna, Italy. [5]Department of Biomedical and Neuromotor Sciences, University of Bologna, Bologna, Italy.

Received: 23 September 2021 Accepted: 2 February 2022
Published online: 10 February 2022

References

1. Kousi M, Katsanis N. Genetic modifiers and oligogenic inheritance. Cold Spring Harb Perspect Med. 2015;5:1–22.
2. Rahit KMTH, Tarailo-Graovac M. Genetic modifiers and rare mendelian disease. Genes (Basel). 2020;11
3. Paré G, Mao S, Deng WQ. A machine-learning heuristic to improve gene score prediction of polygenic traits. Sci Rep. 2017;7:1–11.

4. Xu C, Jackson SA. Machine learning and complex biological data. Genome Biol. 2019;20:76.
5. Laing C, et al. The application of unsupervised clustering methods to Alzheimer's disease. Front Comput Neurosci. 2019;1:31.
6. Bersanelli M, Mosca E, Remondini D, Castellani G, Milanesi L. Network diffusion-based analysis of high-throughput data for the detection of differentially enriched modules. Sci Rep. 2016;6:1–12.
7. Mosca E, et al. Characterization and comparison of gene-centered human interactomes. Brief Bioinform. 2021;2021:1–16.
8. Lopez C, Tucker S, Salameh T, Tucker C. An unsupervised machine learning method for discovering patient clusters based on genetic signatures. J Biomed Inform. 2018;85:30–9.
9. Omta WA, et al. Combining supervised and unsupervised machine learning methods for phenotypic functional genomics. Screening. 2020. https://doi.org/10.1177/24725552091934525,655-664.
10. Libbrecht MW, Stafford Noble W. Machine learning applications in genetics and genomics. Nat Publ Gr. 2015. https://doi.org/10.1038/nrg3920.
11. Pihlstrøm L, Wiethoff S, Houlden H. Genetics of neurodegenerative diseases: an overview. Handbook of clinical neurology, vol. 145. Hoboken: Elsevier; 2018.
12. Jain N, Chen-Plotkin AS. Genetic Modifiers in Neurodegeneration. Curr Genet Med Rep. 2018;6:11–9.
13. Pang SYY, et al. The role of gene variants in the pathogenesis of neurodegenerative disorders as revealed by next generation sequencing studies: a review. Transl Neurodegener. 2017;6:1–11.
14. Cacace R, Sleegers K, Van Broeckhoven C. Molecular genetics of early-onset Alzheimer's disease revisited. Alzheimer's Dementia. 2016. https://doi.org/10.1016/j.jalz.2016.01.012.
15. Poleggi A, et al. Age at onset of genetic (E200K) and sporadic Creutzfeldt–Jakob diseases is modulated by the CYP4X1 gene. J Neurol Neurosurg Psychiatry. 2018;89:1243–9.
16. Hermann P, et al. Biomarkers and diagnostic guidelines for sporadic Creutzfeldt-Jakob disease. Lancet Neurol. 2021;20.
17. Capellari S, Strammiello R, Saverioni D, Kretzschmar H, Parchi P. Genetic Creutzfeldt–Jakob disease and fatal familial insomnia: Insights into phenotypic variability and disease pathogenesis. Acta Neuropathol. 2011;121:21–37.
18. Ladogana A, et al. High incidence of genetic human transmissible spongiform encephalopathies in Italy. Neurology. 2005;64:1592–7.
19. Bellenguez C, et al. Contribution to Alzheimer's disease risk of rare variants in TREM2, SORL1, and ABCA7 in 1779 cases and 1273 controls. Neurobiol Aging. 2017. https://doi.org/10.1016/j.neurobiolaging.2017.07.001.
20. Ridge PG, Mukherjee S, Crane PK, Kauwe JSK, Consortium ADG. Alzheimer's disease: analyzing the missing heritability. PLoS ONE. 2013;8:e79771.
21. Cruchaga C, et al. Polygenic risk score of sporadic late-onset Alzheimer's disease reveals a shared architecture with the familial and early-onset forms. Alzheimer's Dement. 2018;14:205–14.
22. McKhann GM, et al. The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement. 2011;7:263–9.
23. Dubois B, et al. Advancing research diagnostic criteria for Alzheimer's disease: The IWG-2 criteria. Lancet Neurol. 2014. https://doi.org/10.1016/S1474-4422(14)70090-0.
24. Zerr I, et al. Updated clinical diagnostic criteria for sporadic Creutzfeldt–Jakob disease. Brain. 2009;132:2659.
25. Bartoletti-Stella A, et al. Identification of rare genetic variants in Italian patients with dementia by targeted gene sequencing. Neurobiol Aging. 2018;66(180):e23-180.e31.
26. Van Giau V, An SSA, Bagyinszky E, Kim SY. Gene panels and primers for next generation sequencing studies on neurodegenerative disorders. Mol Cell Toxicol. 2015;11:89–143.
27. McKenna A, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.
28. Pedregosa F et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011.
29. Waskom M. Seaborn: statistical data visualization. Seaborn;2012.
30. Plotly Technologies Inc. Collaborative data science, https://plot.ly. Plotly Technologies Inc.;2015.
31. Karczewski KJ, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581:434–43.
32. Rizzu P, et al. High prevalence of mutations in the microtubule-associated protein tau in a population study of frontotemporal dementia in the Netherlands. Am J Hum Genet. 1999. https://doi.org/10.1086/302256.
33. Zabetian CP, et al. Association analysis of MAPT H1 haplotype and subhaplotypes in Parkinson's disease. Ann Neurol. 2007. https://doi.org/10.1002/ana.21157.
34. Package 'HardyWeinberg' Type Package Title Statistical Tests and Graphics for Hardy-Weinberg Equilibrium. 2021; https://doi.org/10.1126/science.28.706.49
35. Huttlin EL, et al. The BioPlex network: a systematic exploration of the human interactome. Cell. 2015. https://doi.org/10.1016/j.cell.2015.06.043.
36. Hein MY, et al. A human interactome in three quantitative dimensions organized by stoichiometries and abundances. Cell. 2015. https://doi.org/10.1016/j.cell.2015.09.053.
37. Wan C, et al. Panorama of ancient metazoan macromolecular complexes. Nature. 2015. https://doi.org/10.1038/nature14877.
38. Rolland T, et al. A proteome-scale map of the human interactome network. Cell. 2014. https://doi.org/10.1016/j.cell.2014.10.050.
39. Luck K, Sheynkman GM, Zhang I, Vidal M. Proteome-scale human interactomics. Trends Biochem Sci. 2017. https://doi.org/10.1016/j.tibs.2017.02.006.
40. Ashburner M, et al. Gene ontology: tool for the unification of biology. Nat Genet. 2000. https://doi.org/10.1038/75556.
41. The Gene Ontology, C. *et al.* The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res. 2019; https://doi.org/10.17863/CAM.36439.
42. Canu E, et al. H1 haplotype of the MAPT gene is associated with lower regional gray matter volume in healthy carriers. Eur J Hum Genet. 2009. https://doi.org/10.1038/ejhg.2008.185.
43. Testi S, et al. Mutational and haplotype map of NOTCH3 in a cohort of Italian patients with cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy (CADASIL). J Neurol Sci. 2012;319:37–41.
44. Kopanos C, et al. VarSome: the human genomic variant search engine. Bioinformatics. 2019;35:1978–80.
45. Amberger JS, Bocchini CA, Scott AF, Hamosh A. OMIM.org: leveraging knowledge across phenotype-gene relationships. Nucleic Acids Res. 2019;47:D1038–43.
46. Landrum MJ, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018;46:D1062–7.
47. Stenson PD, et al. Human Gene Mutation Database (HGMD): 2003 update. Hum Mutat. 2003;21:577–81.
48. Richards S, Aziz N, Bale S, Bick D, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology Sue. Genet Med. 2015;17:405–24.
49. Lacour M, et al. Causative mutations and genetic risk factors in sporadic early onset Alzheimer's disease before 51 years. J Alzheimer's Dis. 2019. https://doi.org/10.3233/JAD-190193.
50. Lambert JC, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat Genet. 2013;45:1452–8.
51. Genin E, et al. APOE and Alzheimer disease: a major gene with semi-dominant inheritance. Mol Psychiatry. 2011. https://doi.org/10.1038/mp.2011.52.
52. Cochran JN et al. Genome sequencing for early-onset or atypical dementia: high diagnostic yield and frequent observation of multiple contributory alleles. Cold Spring Harb Mol Case Stud. 2019;5.
53. Patel D et al. Association of rare coding mutations with alzheimer disease and other dementias among adults of European Ancestry. JAMA Netw Open 2019;2.
54. Skipper L et al. Linkage disequilibrium and association of MAPT H1 in Parkinson disease. Am J Hum Genet. 2004;75.
55. Pittman AM, et al. The structure of the tau haplotype in controls and in progressive supranuclear palsy. Hum Mol Genet. 2004. https://doi.org/10.1093/hmg/ddh138.

56. Caffrey TM, Wade-Martins R. Functional MAPT haplotypes: bridging the gap between genotype and neuropathology. Neurobiol Dis. 2007;27:1–10.
57. Santa-Maria I, et al. The MAPT H1 haplotype is associated with tangle-predominant dementia. Acta Neuropathol. 2012;124:693–704.
58. Sassi C, et al. Mendelian adult-onset leukodystrophy genes in Alzheimer's disease: critical influence of CSF1R and NOTCH3. Neurobiol Aging. 2018. https://doi.org/10.1016/j.neurobiolaging.2018.01.015.
59. Giau VV, et al. Genetic analyses of early-onset Alzheimer's disease using next generation sequencing. Sci Rep. 2019;9:1–10.
60. Myers AJ, Kaleem MA, Marlowe L, Pittman AM. The H1c haplotype at the MAPT locus is associated with Alzheimer 's disease. Hum Mol Genet. 2005. https://doi.org/10.1093/hmg/ddi241.
61. Sánchez-Juan P et al. The MAPT H1 haplotype is a risk factor for Alzheimer's disease in APOE ε4 non-carriers. Front Aging Neurosci. 2019;11.
62. Minikel EV, et al. Quantifying prion disease penetrance using large population control cohorts. Sci Transl Med. 2016. https://doi.org/10.1126/scitranslmed.aad5169.
63. Machado do Nascimento P, Gomes Medeiros I, Maia Falcão R, Stransky B, Santana E, de Souza J. A decision tree to improve identification of pathogenic mutations in clinical practice. BMC Med Inform Decision Mak. 2020. https://doi.org/10.1186/s12911-020-1060-0.
64. Karch CM, Goate AM. Alzheimer's disease risk genes and mechanisms of disease pathogenesis. Biol Psychiat. 2015. https://doi.org/10.1016/j.biopsych.2014.05.006.
65. Wallon D, et al. APP, PSEN1, and PSEN2 mutations in early-onset Alzheimer disease: a genetic screening study of familial and sporadic cases. PLoS Med. 2017;14:1–16.
66. Zheng Q, et al. Dysregulation of ubiquitin-proteasome system in neurodegenerative diseases. Front Aging Neurosci. 2016;8:303.
67. Jones E, et al. Identification of novel risk loci and causal insights for sporadic Creutzfeldt–Jakob disease: a genome-wide association study. Lancet Neurol. 2020;19:840–8.
68. Bartoletti-Stella A, et al. Analysis of RNA expression profiles identifies dysregulated vesicle trafficking pathways in Creutzfeldt-Jakob Disease. Mol Neurobiol. 2019;56:5009–24.
69. Sorce S, et al. Genome-wide transcriptomics identifies an early preclinical signature of prion infection. BioRxiv. 2020. https://doi.org/10.1101/2020.01.10.901637.
70. Poggiolini, I., Saverioni, D. & Parchi, P. Prion protein misfolding, strains, and neurotoxicity: an update from studies on mammalian prions. Int J Cell Biol. 2013;2013.
71. Sazonovs A, Barrett JC. Rare-variant studies to complement genome-wide association studies. Annu Rev Genomics Hum Genet. 2018;19:97–112.
72. Pilla E, Schneider K, Bertolotti A. Coping with protein quality control failure. Annu Rev Cell Dev Biol. 2017;33:439–65.
73. Labzin LI, Heneka MT, Latz E. Innate immunity and neurodegeneration. Annu Rev Med. 2018;69:437–49.
74. Abu-Rumeileh S, et al. CSF biomarkers of neuroinflammation in distinct forms and subtypes of neurodegenerative dementia. Alzheimer's Res Ther. 2019;12:2.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

BMC