Journal of the American Statistical Association
December 2025 • Volume 120 • Number 552

Taylor & Francis
Taylor & Francis Group

Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: www.tandfonline.com/journals/uasa20

SurvSTAAR: A powerful statistical framework for rare variant analysis of time-to-event traits in large-scale whole-genome sequencing studies

Yidan Cui, Shiyang Ma, Yuxin Yuan, Nengjie Zhu, Haifeng Chen, Ting Wei, Zilin Li, Xihao Li & Zhangsheng Yu

To cite this article: Yidan Cui, Shiyang Ma, Yuxin Yuan, Nengjie Zhu, Haifeng Chen, Ting Wei, Zilin Li, Xihao Li & Zhangsheng Yu (05 Jan 2026): SurvSTAAR: A powerful statistical framework for rare variant analysis of time-to-event traits in large-scale whole-genome sequencing studies, Journal of the American Statistical Association, DOI: 10.1080/01621459.2025.2606388

To link to this article: https://doi.org/10.1080/01621459.2025.2606388

SurvSTAAR: A powerful statistical framework for rare variant analysis of time-to-event traits in large-scale whole-genome sequencing studies

Yidan Cui[a,†], Shiyang Ma[b,c,†], Yuxin Yuan[d], Nengjie Zhu[c], Haifeng Chen[a], Ting Wei[a,e], Zilin Li[d,*], Xihao Li[f,g,*], Zhangsheng Yu[a,b,*]

[a]Department of Bioinformatics and Biostatistics, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, Chin

[b]Institute of Clinical Medicine, Shanghai Jiao Tong University School of Medicine, Shanghai, China

[c]School of Mathematical Sciences, Shanghai Jiao Tong University, Shanghai, China

[d]School of Mathematics and Statistics and KLAS, Northeast Normal University, Jilin, China

[e]Intelligent Medicine Original Medical Technology (Shanghai) Co., Ltd., Shanghai, China

[f]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[g]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[†]These authors contributed equally: Yidan Cui, Shiyang Ma.

*Correspondence should be addressed to Z. Li (lizl@nenu.edu.cn), X. Li (xihaoli@unc.edu), and Z. Yu (yuzhangsheng@sjtu.edu.cn).

Abstract

The increasing availability of large-scale, population-based whole-genome sequencing (WGS) data enables comprehensive analyses of rare genetic variants, which are crucial for unraveling the genetic mechanisms underlying complex traits and diseases. Time-to-event traits offer the advantage of capturing both diagnosis status and timing, facilitating the identification of genetic variants associated with age of onset, disease progression, and lifespan. However, existing methods primarily focus on quantitative and binary traits, which do not leverage censored time information and have limitations in detecting rare variants associated with disease progression. Here we propose SurvSTAAR, a powerful and comprehensive statistical framework for time-to-event traits in large-scale WGS studies, offering a computationally scalable analytical pipeline for analyzing rare coding and noncoding variants. SurvSTAAR accounts for sample relatedness, population structure, heavily censored traits, and further empowers rare variant association analysis by incorporating functional annotations. We applied SurvSTAAR to analyze the time-to-event trait of Alzheimer's disease (AD) in 458,773 related samples from the UK Biobank WGS data. We identified putatively novel associations with AD in both coding and noncoding regions and further explored their potential role in disease progression by assessing their effects on protein function, including amino acid changes, structural modifications, and domain disruptions.

Keywords: time-to-event; rare variant analysis; whole genome sequencing; functional annotation; Alzheimer's disease

1 Introduction

Alzheimer's disease (AD) is a neurodegenerative disorder with a strong genetic basis and imposes a substantial public health burden worldwide (Nichols et al., 2022). Facilitated by advancements in next-generation sequencing technologies and the decreasing costs of whole-genome sequencing (WGS), the field has recently seen a paradigm shift toward rare variant (RV; minor allele frequency (MAF) < 1%) analysis (De Deyn and Sleegers, 2025; Povysil et al., 2019).

The emergence of large-scale cohorts, such as the UK Biobank (UKB) (Bycroft et al., 2018; Carss et al., 2025), the All of Us Research Program (Bick et al., 2024; The All of Us Research Program Investigators, 2019), and the Trans-Omics for Precision Medicine Program (TOPMed) (Taliun et al., 2021), provide extensive genomic data combined with comprehensive electronic health records (EHRs). These resources enable analysis of time-to-event phenotypes and offer opportunities to investigate the genetic mechanisms of AD initiation and progression. However, existing methods for time-to-event genetic association analysis are mainly restricted to single variant tests for common variants (Bi et al., 2020; Dey et al., 2022; He and Kulminski, 2020; Ojavee et al., 2021; Pedersen et al., 2023; Rizvi et al., 2019), resulting in limited power to detect RV associations.

To address the limitation in detecting RV associations, variant-set tests have been widely adopted for quantitative and binary traits, aggregating information to examine the joint effects of multiple variants within a set (Lee et al., 2014; Li and Leal, 2008; Liu et al., 2019; Madsen and Browning, 2009; Morgenthaler and Thilly, 2007; Morris and Zeggini, 2010; Wu et al., 2011). Existing methods, such as Cox SKAT (Chen et al., 2014), extend variant-set tests for time-to-event traits. However, they do not scale to biobank-level WGS data and cannot adequately account for sample relatedness or heavy censoring, both of which are common in large-scale cohorts. Furthermore, existing methods do not integrate variant functional annotations, leading to limited interpretability and reduced power. Although the STAAR framework (Li et al., 2020, 2022) dynamically incorporates multiple functional annotations to empower RV association analysis, it is specifically tailored for quantitative and binary traits.

Here, we introduce SurvSTAAR, a powerful statistical framework and computationally scalable analytical pipeline that enables functionally informed genetic association analysis for time-to-event traits in large-scale WGS studies (**Figure 1**). It has several features. First, SurvSTAAR accounts for sample relatedness and population structure by incorporating polygenic score (Mbatchou et al., 2021) and adjusting for ancestral principal components (PCs) in the Cox proportional hazards (PH) model (Cox, 1972). SurvSTAAR further addresses challenges posed by heavily censored phenotypes using Saddlepoint approximation (SPA) (Bi et al., 2020; Dey et al., 2017; Feuerverger, 1989) to calibrate *p*-values. Second, SurvSTAAR incorporate multiple variant functional annotations, which capture various aspects of variant function, as combined weights in set-based tests, thereby enhancing the power of RV association studies. Third, we additionally developed a corresponding analysis pipeline of SurvSTAAR, enabling scalable, flexible, and streamlined functionally informed time-to-event-trait genetic association analysis for biobank-scale sequencing data, particularly in the noncoding genome.

This paper is structured as follows. Section 2 describes the UKB dataset and key characteristics of the AD onset phenotype, which motivate our methodological development. Section 3 introduces the SurvSTAAR framework and outlines its key methodological components. Section 4 presents the application of SurvSTAAR to coding and noncoding RVs in the UKB WGS data for AD onset, followed by downstream analyses to explore molecular mechanisms underlying AD progression. Section 5 presents extensive simulation studies demonstrating that SurvSTAAR maintains valid type I error control and achieves substantial power gains. Section 6 concludes with a discussion.

2 Data Description

The UKB is a large-scale prospective cohort study that recruited approximately half a million participants aged 40-69 across the United Kingdom between 2006 and 2010, collecting extensive phenotypic, genomic, and health-related data (Sudlow et al., 2015). We utilized WGS data from UKB to investigate RVs associated with the AD progression. We applied quality control procedures consistent with previous UKB WGS studies (Carss et al., 2025; Halldorsson et al., 2022; Li et al., 2025), with detailed genomic data quality control are provided in the **Supplementary Notes**.

After applying sample- and variant-level quality control, approximately 1.1 billion variants were observed among 458,773 individuals. Notably, more than 90% of these WGS variants are rare and are predominantly located in noncoding regions. Prior UKB WGS-based heritability studies shows that RVs contribute approximately 20% of pedigree-based heritability on average, with only 21% attributable to coding regions and the remaining 79% arising from noncoding regions (Wainschtein et al., 2025). These findings underscore the critical need for statistical methods that can effectively detect RV associations across both coding and noncoding regions in large-scale WGS datasets.

The study population displays three key characteristics. First, more than 80% of disease phenotypes are highly imbalanced (Zhou et al., 2018), a pattern analogous to heavy censoring in time-to-event settings. Second, the cohort exhibits substantial sample relatedness, with more than 30% of participants having at least one relative up to the third degree. Third, the cohort exhibits complex population structure, with more than 30,000 participants not identified as non-Finnish Europeans based on genetic ancestry (Carss et al., 2025). These population and genomic features collectively motivate the development of SurvSTAAR, a statistical framework designed to analyze both coding and noncoding RVs for time-to-event traits.

In the context of AD, although considerable advances has been made in identifying genetic risk factors, most studies treat AD as a binary outcome, overlooking the dynamics of disease progression (Jack et al., 2024). The UKB provides precise diagnosis dates, enabling the analysis of AD progression through the construction of high-quality time-to-event AD phenotypes (detailed in **Supplementary Notes**), which yielded 3,953 AD events. As RVs are increasingly recognized as being more directly linked to trait biology (Spence et al., 2025), the UKB WGS data facilitate the identification of AD progression-related RVs and the investigation of their potential molecular mechanisms across the whole genome. Applying SurvSTAAR to this dataset enables the discovery of putatively novel RVs associated with AD progression, capturing signals from both coding and noncoding regions.

3 Modeling Framework

3.1 Notation and Statistical Model

Suppose there are n subjects with M variants sequenced across the whole genome. For subject i, let T_{ci} denote the potential censoring time, and T_{fi} the potential failure time. The observed event time for subject i is given by $T_i = \min(T_{fi}, T_{ci})$, and the event status is indicated by $\Delta_i = I(T_{fi} \leq T_{ci})$, where $I(\cdot)$ is the indicator function. Thus, the observed time-to-event phenotype for subject i is $(T_i = t_i,\ \Delta_i = \delta_i)$. For a genetic set of p variants, let $\boldsymbol{G}$ represent the $n\ \times p$ matrix, where each element $G_{ij} \in \{0, 1, 2\}$ represents the allele count for individual i at the j-th marker. Similarly, let $\boldsymbol{X}$ denote the $n \times q$ matrix of covariates, such as sex, birth year, and ancestral PCs. We consider Cox PH model $\lambda_i(t;\ \boldsymbol{X_i}, \boldsymbol{G_i}, \widehat{\omega}_i) = \lambda_0(t) \exp(\boldsymbol{X_i}^\top \boldsymbol{\alpha} + \boldsymbol{G_i}^\top \boldsymbol{\beta} + \widehat{\omega}_i)$, where $\lambda_0(t)$ represents the baseline hazard function, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ represent the covariate and genetic effect coefficients, respectively. The term $\widehat{\omega}_i$, representing individual-level polygenic scores, is included as an offset in the null model to account for sample relatedness due to unobserved familial structure, which could otherwise lead to inflated type I error rates (Xu et al., 2025). Prior to null model fitting, $\widehat{\omega}_i$ is estimated using REGENIE v3.5 (Mbatchou et al., 2021), which performs genome-wide two-level stacked ridge regression using logistic regression. The event indicator $\boldsymbol{\delta}$ is treated as a binary outcome in this step.

3.2 Null Model Specification

The score test can significantly reduce the computational burden in WGS data because the null model remains consistent across all genetic variants, requiring it to be fitted only once for the entire genome-wide analysis. To conduct the score test for the j-th variant in this variant set, we test the null hypothesis $H_{0j}: \beta_j = 0$, under which we fit the null Cox PH model $\lambda_i(t; \boldsymbol{X_i}, \widehat{\omega}_i) = \lambda_0(t) \exp(\boldsymbol{X_i}^\top \boldsymbol{\alpha} + \widehat{\omega}_i)$ and estimate the corresponding martingale residuals R_i using the R package *survival* (Terry M. Therneau, 2024; Terry M. Therneau and Patricia M. Grambsch, 2000). To address the tied survival times, we use Breslow's approximation (Breslow, 1972; Lin, 2007), with details in the **Supplementary Notes**.

3.3 Association Testing Framework

3.3.1 Score-Based Test for Single-Variant Analysis

To perform association analysis for each variant, we construct the score statistic $S = \sum_{i=1}^{n} G_i R_i$, where G_i is the genetic variant for the i-th subject and R_i is the martingale residual. The asymptotic variance is estimated as $\widehat{Var}(S) = \boldsymbol{G}^\top \boldsymbol{V} \boldsymbol{G} - \boldsymbol{G}^\top \boldsymbol{V} \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{V} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{V} \boldsymbol{G}$, with $\boldsymbol{V}$ detailed in the **Supplementary Notes**. The adjusted score test statistic $T_{adj} = S / \sqrt{\widehat{Var}(S)}$ is assumed to follow a standard normal distribution under $H_0: \beta = 0$, with the p-value derived from $T_{adj}^2 \sim \chi_1^2$. However, heavily censored time-to-event phenotypes can lead to skewed martingale residuals, causing a right-skewed null distribution of T_{adj} and inflated type I error rates.

To address this challenge, we adopt the SPA method, which offers greater accuracy over the normal approximation by using the entire cumulative generating function (CGF) to approximate the null distribution of score statistics. However, in the Cox PH model, the null distribution of score statistic S is complex, and its theoretical CGF does not have a closed-form expression. Therefore, we follow the empirical SPA approach (Bi et al., 2020; Feuerverger, 1989) to approximate the CGF and define the centered covariate-adjusted genotypes as $\widetilde{\boldsymbol{G}} = \boldsymbol{G} - \widetilde{\boldsymbol{X}} \left(\widetilde{\boldsymbol{X}}^\top \widetilde{\boldsymbol{X}}\right)^{-1} \widetilde{\boldsymbol{X}}^\top \boldsymbol{G}$, where $\widetilde{\boldsymbol{X}} = [\boldsymbol{1_n}, \boldsymbol{X}]$. We then replace the moment generating function (MGF) of martingale residual, $M(\xi) = E(e^{\xi R})$, with its sample version, $M_n(\xi) = \frac{1}{n} \sum_{i=1}^{n} e^{\xi R_i}$. The corresponding empirical CGF is defined as $K_n(\xi) = \log M_n(\xi)$. The distribution of S can be approximated as $\Pr(S < s) \approx \Phi\left\{w + \frac{1}{w} \cdot \log \frac{v}{w}\right\}$, where Φ is the standard normal cumulative distribution function. The association p-value between a single variant and the time-to-event trait is computed as $p = \Pr(S < -|s|) + \Pr(S > |s|)$, where s is the observed score statistic, and the definition of w and v are provided in the **Supplementary Notes**.

3.3.2 Set-Based Association Test for Rare Variant Analysis

Several score-based aggregation methods are commonly used to assess the association between sets of RVs and phenotypes, including the Burden test (Li and Leal, 2008; Madsen and Browning, 2009;

Morgenthaler and Thilly, 2007; Morris and Zeggini, 2010), the sequence kernel association test (SKAT) (Wu et al., 2011), and the aggregated Cauchy association test (ACAT-V) (Liu et al., 2019). These methods differ in their modeling assumptions and weighting strategies, and have been widely applied in RV association studies. Further details on these set-based methods are provided in the **Supplementary Notes**.

Building upon these conventional methods, SurvSTAAR extends set-based testing strategies to analyze time-to-event phenotypes, resulting in SurvBurden, SurvSKAT, and SurvACAT-V. To address heavy censoring, SurvSTAAR further incorporates SPA to approximate set-based p-values. Let $\tilde{p}$ denote the SPA-adjusted single-variant p-value. For the SurvBurden test, the aggregate score $S_{\text{SurvBurden}}$ is a linear combination of single-variant score statistics. Therefore, SPA can be directly applied to approximate its null distribution and compute the calibrated p-value $\tilde{p}_{\text{SurvBurden}}$. For the SurvACAT-V test, $\tilde{p}_j$ are combined following the conventional ACAT-V strategy, which involves a Cauchy transformation and weighted aggregation.

For the SurvSKAT test, the test statistic follows a mixture of chi-square distributions under the null hypothesis, making SPA inapplicable. To address this, we propose a novel approximation within the SurvSTAAR framework to obtain calibrated p-values under heavy censoring. The inverse standard normal transformation is applied to obtain $\tilde{z}_j = \Phi^{-1}(\tilde{p}_j)$, where $\Phi^{-1}(\cdot)$ denotes the quantile function of the standard normal distribution. Let Σ denote the covariance matrix of the original score test statistics, where $\Sigma_{m,n}$ represents its (m, n)-th element. We make the following assumptions:

- $(\tilde{z}_1, \dots, \tilde{z}_j) \sim MVN(0, \tilde{\Sigma})$, where for diagonal elements $(m = n)$, $\tilde{\Sigma}_{m,n} = 1$, and for off-diagonal elements $(m \neq n)$, $\tilde{\Sigma}_{m,n} = cov(\tilde{z}_m, \tilde{z}_n)$.
- For RVs, we assume $cov(\tilde{z}_m, \tilde{z}_n) \approx cov(S_m, S_n) = \Sigma_{m,n}$, where S_m and S_n are individual score test statistics for different variants.

Thus, the SurvSKAT test statistic can be expressed as $Q_{\text{SurvSKAT}} = \sum_j w_j^2 \Sigma_{jj} \tilde{z}_j^2$ and $Q_{\text{SurvSKAT}} \sim \sum_j \lambda_j \chi^2_{1,j}$, where λ_j are the eigenvalues of $\tilde{\Sigma}$, and $\chi^2_{1,j}$ are independent chi-square distributions with 1 degree of freedom. To obtain the p-value, we use Davies' exact method (Davies, 1980) by default. If Davies' p-value is 0 or fails, we switch to Liu's method (Liu et al., 2009). To handle extremely RVs (e.g., minor allele count (MAC) ≤ 20), we first aggregate them using SurvBurden weights, following a strategy similar to the ACAT-V test (see **Supplementary Notes**). To ensure the combined variant has the same scale as other variants, we apply L_2 standardization to those SurvBurden weights.

After establishing SurvBurden, SurvSKAT, and SurvACAT-V tests for heavily censored time-to-event phenotypes, we further enhance our framework by integrating functional annotations. Conventional methods directly assume effect sizes β_j, overlooking the causal relationships between RVs and phenotypes. In contrast, SurvSTAAR integrates functional annotations into set-based analysis to assess the probability weight of a variant being causal to the phenotype.

Specifically, the weights within SurvSTAAR combines two components: the causal probability $\hat{\pi}_{jk}$, estimated from the k-th functional annotation, and the weight w_{jl} derived from the l-th Beta distribution Beta (a_{l1}, a_{l2}). Here, $k = 0, \dots, K$, and $l = 1, \dots, L$, resulting in $L \times (K + 1)$ distinct weights. Specifically, within the SurvSTAAR framework, the test statistics for SurvBurden, SurvSKAT, and SurvACAT-V are defined as:

$$Q_{\mathrm{SurvBurden},k,l} = \left(\sum_{j=1}^{p} \hat{\pi}_{jk}\, w_{jl} S_j\right)^2,$$

$$Q_{\mathrm{SurvSKAT},k,l} = \sum_{j=1}^{p} \hat{\pi}_{jk}\, w_{jl}^2 S_j^2,$$

$$Q_{\mathrm{SurvACAT-V},k,l} = \sum_{j'=0}^{s} \hat{\pi}_{j'k}\, w_{A,j',l}^2 \tan\{(0.5 - p_{j'})\pi\},$$

with corresponding p-values $p_{\mathrm{SurvBurden},k,l}$, $p_{\mathrm{SurvSKAT},k,l}$, and $p_{\mathrm{SurvACAT-V},k,l}$, respectively. The weight $w_{A,j'}$ used in the SurvACAT-V test is analogous to that in the ACAT-V test. Further details on causal probability estimation and weight construction procedure in SurvSTAAR are provided in the **Supplementary Notes**.

We define three set-based tests within SurvSTAAR framework: SurvSTAAR-Burden, SurvSTAAR-SKAT, and SurvSTAAR-ACAT-V. The general form of the SurvSTAAR-test statistic is given by:

$$T_{\mathrm{SurvSTAAR-test}} = \sum_{l=1}^{L} \sum_{k=0}^{K} \frac{\tan\{(0.5 - p_{\mathrm{test},k,l})\pi\}}{L \times (K+1)},$$

where test ∈ {SurvBurden, SurvSKAT, and SurvACAT-V}. Then we have the corresponding p-value $p_{\mathrm{SurvSTAAR-test}} \approx 1/2 - \arctan(T_{\mathrm{SurvSTAAR-test}})/\pi$. The SurvSTAAR-O test for the time-to-event trait can be computed as: $\mathrm{SurvSTAAR} - \mathrm{O} = \frac{1}{3}\{\tan[(0.5 - p_{\mathrm{SurvSTAAR-B}})\pi] + \tan[(0.5 - p_{\mathrm{SurvSTAAR-S}})\pi] + \tan[(0.5 - p_{\mathrm{SurvSTAAR-A}})\pi]\}$, with the corresponding p-value $p_{\mathrm{SurvSTAAR-O}} \approx 1/2 - \arctan(T_{\mathrm{SurvSTAAR-O}})/\pi$.

4 Application to Alzheimer's Disease Progression in UKB WGS Data

To demonstrate the practical utility of SurvSTAAR in biobank-scale cohorts, we apply it to the AD progression phenotype and WGS data described in Section 2.

4.1 Conditional Analysis Approach

We conducted conditional analysis to identify novel RV associations by adjusting for known phenotype-related variants. We downloaded all known variants from the GWAS catalog (Sollis et al., 2023), using trait labels *Alzheimer disease* (EFO ID: MONDO_0004975) and *age of onset of Alzheimer disease* (EFO ID: OBA_2001000). A step-by-step algorithm, analogous to STAARpipeline (Li et al., 2022), was employed to select a subset of independent variants representing all known variants on each chromosome, with further details provided in the **Supplementary Notes**.

4.2 Gene-Based Analysis of Coding and Non-Coding Rare Variants

We utilized SurvSTAAR-O test to perform gene-centric analysis of coding and noncoding RVs (MAF < 0.01) based on functional categories. We adjusted for sex, birth year, the first ten genetic principal components accounting for population stratification, and the polygenic score accounting for sample relatedness.

We defined five functional categories to aggregate protein-coding RVs, including putative loss-of-function (pLoF; stop gain, stop loss and splice); missense; disruptive missense; synonymous; and the combined pLoF and disruptive missense (pLoF+D). The pLoF, missense, and synonymous RVs were defined by GENCODE Variant Effect Predictor (VEP) categories (Frankish et al., 2019; Harrow et al., 2012). The disruptive variants were further defined by MetaSVM (Dong et al., 2015), which measures the deleteriousness of missense mutations. For each functional category, we incorporate 9 annotation principal components (aPCs) (Li et al., 2020; Zhou et al., 2023), and CADD (Kircher et al., 2014), LINSIGHT (Huang et al., 2017), FATHMM-XF (Rogers et al., 2018), and MetaSVM (Dong et al., 2015) (for missense RVs only) along with two MAF-based weights to enhance power (**Supplementary Table 1**). The overall distribution of SurvSTAAR-O *p*-values was well calibrated for the time-to-event trait analysis of coding RVs. To determine significance, we applied a Bonferroni adjusted genome-wide significant threshold of $\alpha = 5.00\times10^{-7}$ (0.05/[20,000×5]), accounting for five functional coding categories across protein-coding genes. SurvSTAAR-O identified seven genome-wide significant associations from unconditional analyses. After conditioning on the previously reported AD-related variants (ARUK Consortium et al., 2017; Bellenguez et al., 2022; He et al., 2021; Jansen et al., 2019; Moreno-Grau et al., 2019; Schwartzentruber et al., 2021; Wightman et al., 2021), five out of seven associations remained significant at the Bonferroni-adjusted threshold ($\alpha = 7.14\times10^{-3}$ (0.05/7)), including disruptive missense and pLoF+D RVs of *SORL1*, and missense RVs of *TOMM40*, *APOE*, and *CLPTM1* (**Table 1**, **Figure 2a-b**, **Supplementary Table 2**).

For gene-centric noncoding RV analysis, eight functional categories were established, including upstream region; downstream region; untranslated regions (UTRs); promoter RVs overlaid with Cap Analysis of Gene Expression (CAGE) sites; promoter RVs overlaid with DNase hypersensitivity (DHS) sites; enhancer RVs overlaid with CAGE sites; enhancer RVs overlaid with DHS sites; noncoding RNA (ncRNA) RVs. The promoter RVs were defined as those located within a ±3-kilobase (kb) window around transcription start sites, while the enhancer RVs were defined as those located in GeneHancer-predicted regions that overlap with CAGE or DHS sites (Fishilevich et al., 2017; The ENCODE Project Consortium, 2012; The ENCODE Project Consortium et al., 2020; The FANTOM Consortium et al., 2014; The FANTOM Consortium and the RIKEN PMI and CLST (DGT), 2014). We defined the UTR, upstream, downstream and ncRNA RVs by GENCODE VEP categories (Frankish et al., 2019; Harrow et al., 2012). For the UTR mask, we included RVs in both 5′ and 3′ UTR regions. We considered the protein-coding gene for the first seven categories provided by Ensembl (Kinsella et al., 2011) and the ncRNA genes provided by GENCODE (Frankish et al., 2019; Harrow et al., 2012). Using a Bonferroni-corrected significance threshold, we set $\alpha = 3.57\times10^{-7}$ (0.05/[20,000×7]) for seven functional noncoding categories across protein-coding genes, and $\alpha = 2.50\times10^{-6}$ (0.05/20,000) for ncRNA genes. With a well-calibrated overall distribution of *p*-values, SurvSTAAR-O identified thirteen genome-wide significant associations from unconditional analyses. After conditioning on the previously reported AD-related variants (ARUK Consortium et al., 2017; Bellenguez et al., 2022; He et al., 2021; Jansen et al., 2019; Moreno-Grau et al., 2019; Schwartzentruber et al., 2021; Wightman et al., 2021), two out of thirteen associations remained significant at the Bonferroni-corrected threshold of $\alpha = 3.85\times10^{-3}$ (0.05/13), including UTR of *BCAM* and enhancer CAGE of *NECTIN2* (**Table 2**, **Figure 2c-d**, **Supplementary Table 3**). No significant signal was identified in ncRNA RVs analysis (**Figure 2e-f**).

4.3 Impact of Rare Variant Mutations on Protein Structure and Function

Given that *SORL1* exhibited the most significant result in the conditional analysis (**Table 1**), we extended our investigation from RV analysis to protein structure analysis, focusing on the potential impact of RV mutations on the structure and function of the corresponding protein. Sortilin-related receptor (SORL1, also known as sorLA or LR11) is a key regulator of amyloid precursor protein (APP) trafficking and recycling, playing a crucial role in preventing the amyloidogenic cleavage of APP into amyloid β-peptide (Aβ), the main component of neuritic plaques in AD (Andersen et al., 2005; Zhang et al., 2023) (**Figure 3a**).

To assess the structural impact of RVs in *SORL1*, we considered all missense mutations within the functional mask, excluding pLoF RVs, due to their substantially greater impact on protein structure. For amino acid positions with multiple missense mutations, we selected the most pathogenic variant based on AlphaMissense predictions (Cheng et al., 2023). The wild-type SORL1 structure was obtained from the AlphaFold Protein Structure Database, and the mutant structure was predicted using AlphaFold3 (Abramson et al., 2024). Signal peptides were predicted using SignalP 6.0 (Teufel et al., 2022).

We used the "*align*" function in PyMOL to compare the structural differences between wild-type and missense-mutant SORL1 proteins. The analysis revealed significant structural alterations, with a root-mean-square deviation (RMSD) of 42.86 Å, indicating substantial differences between the two structures (**Figure 3b-c**). In the signal peptide domain (residues 1-28), five amino acids are affected by seven RV missense mutations. Notably, the p.M1T mutation is highly prevalent in AD patients (mutation frequency: 1.26×10^{-3} in AD cases, 1.54×10^{-5} in censored individuals). Methionine (M), the first amino acid, serves as the start codon and is hydrophobic, whereas a mutation to threonine (T) is hydrophilic, potentially disrupting the hydrophobic core of the signal peptide and affecting protein localization. Beyond the signal peptide, RV missense mutations are primarily concentrated in the YWTD and CR regions (**Figure 3b-c**). These domains are essential for APP binding (Andersen et al., 2023; Fazeli et al., 2024; Holstege et al., 2023; Mehmedbasic et al., 2015), and mutations in these regions may disrupt SORL1-APP interactions, leading to aberrant APP trafficking and increased Aβ production, thereby driving AD progression. Additionally, pLoF variants have even more devastating effects on SORL1 structure, potentially causing complete loss of APP binding or severe defects in endosomal sorting, ultimately accelerating AD onset and progression.

5 Simulation Studies

To evaluate type I error control and empirical power of SurvSTAAR, we conducted simulations under various scenarios, including heavily censored traits and settings with a high proportion of related individuals. We generated genotypes for 10,000 independent individuals across 100 distinct 1-Mb regions using the coalescent model (Schaffner et al., 2005) calibrated to mimic the linkage disequilibrium patterns of the European population. In all scenarios, RVs with MAF < 1% were aggregated for testing.

For each subject i, we generated a censoring time T_{ci} and an underlying failure time T_{fi}, defining the observed time-to-event phenotype as $T_i = \min(T_{fi}, T_{ci})$ with the status indicator $\delta_i = I(T_{fi} \leq T_{ci})$. The censoring time T_{ci} followed an exponential distribution with mean $1/\lambda_c$. The underlying failure time T_{fi} was generated using a Cox PH model with a Weibull baseline hazard function as $T_{fi} = -\log U_i/\lambda \exp(\eta_i)$, where $U_i \sim \text{uniform}(0,1)$ and η_i is the linear predictor.

5.1 Type I Error Evaluation

To evaluate the type I error control, we conducted simulations using randomly selected 5-kb segments from pre-generated 1-Mb regions. Under the null hypothesis of no genetic effects, we set $\eta_i = 0.5X_{1i} + 0.5X_{2i}$, where $X_1 \sim \text{Bernoulli}(0.5)$ and $X_2 \sim N(0, 1)$. The event-to-censoring ratio was set to approximately 1:9. Empirical type I error rates were assessed using 10^8 replicates at significance levels $\alpha = 10^{-4}$, 10^{-5}, and 10^{-6}. The simulation results demonstrate that all tests within SurvSTAAR framework generally maintain robust control of type I error rates, with values closely matching the nominal α levels (**Supplementary Table 11**).

We further conducted simulations based on UKB imputed data to evaluate the control of type I error rates under sample relatedness, focusing on coding RVs (with quality control details provided in the **Supplementary Notes**). To better reflect the characteristics of real-data analysis, we randomly selected 100,000 samples, including 50,000 individuals having relatives up to the third degree. Two individuals were considered related if their UKB-reported kinship coefficient exceeded 0.044. The phenotype was generated following the same procedure as in the previous simulations, with the event-to-censoring ratio set to 1:99. We conducted 7,500 replicates, resulting in approximately 10^8 set-based tests. **Supplementary Table 12** demonstrates that all tests within SurvSTAAR maintain well-controlled type I error rates.

5.2 Power Evaluation

In each simulation replicate, we randomly selected 5-kb segments from the 1-Mb regions and assigned ten annotations (A1-A10) to each variant from $N(0,1)$. The linear predictor was defined as $\eta_i = 0.5X_{1i} + 0.5X_{2i} + \sum \boldsymbol{G}_i^\top \boldsymbol{\beta}$, where X_{1i} and X_{2i} are covariates, $\boldsymbol{G}_i$ is the vector of causal variants, and $\beta_j = c_0 |\log_{10} \mathrm{MAF}_j|$ represents the effect size for variant j, with the constant c_0 fixed at 0.5. To evaluate power under varying signal strengths and effect directions, we set the proportion of causal variants within each region to approximately 5%, 15%, and 35%, and randomly assigned 20% or 50% of causal variants with negative effects (details provided in the **Supplementary Notes**).

Empirical power was estimated over 10^4 replicates as the proportion of *p*-values below 10^{-7}, with the event-to-censoring ratio set to 1:9. We compared SurvSTAAR with Cox SKAT (Chen et al., 2014) using survival outcomes, and with STAAR (Li et al., 2020) by treating survival status as a binary trait. Across all scenarios, SurvSTAAR-O consistently outperformed other tests (**Supplementary Figures 1-4**), highlighting the benefits of incorporating functional annotations and harnessing censored event time information in RV analysis.

6 Discussion

This study introduced the SurvSTAAR as a general statistical framework and a comprehensive analytical pipeline for functionally informed RV analysis of time-to-event traits in large-scale WGS studies. SurvSTAAR accounts for population structure, sample relatedness, heavily censored phenotypes, potentially improving power by incorporating both diagnosis time and status of complex disease while harnessing censored event time information. Additionally, SurvSTAAR enables conditional analyses to identify potentially novel RV associations, independent of known variants. Applying SurvSTAAR to UKB WGS data from nearly half a million individuals, we identified five coding and two noncoding conditionally significant RV associations with the time-to-event AD trait.

Among the seven identified RV associations, the association of both pLoF+D RVs, as well as disruptive missense RVs alone in *SORL1*, remained significant under the stringent threshold of 5.00×10^{-7} in both unconditional and conditional analyses. Notably, the *p*-value for the category of pLoF+D RVs was lower than that for disruptive missense RVs alone. The pLoF RVs alone did not reach genome-wide significance at this threshold, suggesting that disruptive missense RVs primarily drive the association, although pLoF RVs also contribute. The most significant result from the conventional variant set tests for the combined RVs in *SORL1* was observed in the SurvBurden test with $P = 4.15\times10^{-10}$. This association was driven by the cumulative effects of multiple missense RVs, rather than any single variant. All RVs with MAC > 20 showed *p*-values > 10^{-3}, with extremely RVs contributing most significantly (**Supplementary Table 4**). The robustness and interpretability of this association were supported by hazard ratio estimates under varying MAF cutoffs of 10^{-2}, 10^{-3} and 10^{-4}, and sensitivity analysis by additionally adjusting of *APOE* ε4 carrier status (**Supplementary Tables 5** and **6**). This association was also replicated in the ADNI cohort (**Supplementary Table 7**; details provided in the **Supplementary Notes**), and consistent signals were observed in the All by All browser based on the All of Us dataset (**Supplementary Table 8**). Further analysis of a binary trait defined by AD status revealed that these functional RVs in *SORL1* exhibit a stronger association in the time-to-event trait than in the binary trait, highlighting their impact on AD progression (**Supplementary Table 9**). The protein structure analysis underscores the pathogenic role of RVs in *SORL1*, suggesting that they contribute to AD progression by disrupting SORL1-APP interactions and promoting Aβ accumulation.

We highlighted the SurvSTAAR-O, the omnibus test that aggregates multiple annotation-weighted set-based tests, including SurvSTAAR-Burden, SurvSTAAR-SKAT, SurvSTAAR-ACAT-V, ensuring robustness under different genetics effect directions within a variant set. SurvSTAAR further provides a flexible and comprehensive analytical pipeline for biobank-scale WGS RV analysis of time-to-event traits through gene-centric analysis using various coding and noncoding functional categories. SurvSTAAR also allows customizable variant set definitions, user-specified functional annotation weights, and adjustable MAF or MAC thresholds for defining RV sets. Beyond its capabilities for RV association analysis in coding and noncoding genome, SurvSTAAR also facilitates single variant analysis of common and low-frequency variants, allowing users to specify thresholds for MAF or MAC, such as MAF ≥ 0.01 or MAC ≥ 20.

There are several limitations to this study. First, the analysis pipeline provides gene-centric coding and noncoding analysis focusing on coding and regularity region, and could be further extended to non-gene-centric analysis for detecting RV associations in intergenic regions by incorporating sliding or dynamic window analysis methods (Li et al., 2020, 2019). Second, it could leverage summary statistics to enhance the power of time-to-event trait analysis through meta-analysis of multiple studies or biobanks(Li et al., 2023). Third, SurvSTAAR has the potential to be adapted for analyzing left-truncated data, competing risks, or joint models.

In summary, we proposed the SurvSTAAR framework, a robust statistical tool for detecting RV associations of time-to-event traits in biobank-scale WGS studies. SurvSTAAR enables researchers to investigate the effects of both coding and noncoding RVs on disease onset in large-scale WGS studies. Available in both offline and cloud computing environments, SurvSTAAR offers a flexible and comprehensive tool for integrating genetic insights on complex disease progression.

7 Genome build

All genome coordinates are given in NCBI GRCh38/UCSC hg38.

8 Data availability

The UKB analyses were conducted using the UKB resource under applications 91486 and 100014. Access to the UKB resource is available via application (https://www.ukbiobank.ac.uk/). The whole-genome individual functional annotation data were assembled from a variety of sources, and the computed annotation PCs are available at the Functional Annotation of Variant-Online Resource (FAVOR) site (https://favor.genohub.org) (Zhou et al., 2023) and the FAVOR database (https://doi.org/10.7910/DVN/1VGTJI) (Zhou et al., 2022).

9 Code availability

Genetics data analysis was conducted in the UKB Research Analysis Platform (RAP, https://ukbiobank.dnanexus.com/). SurvSTAAR is implemented as an open-source R package available at https://github.com/Cui-yd/SurvSTAAR. The analytical pipeline and implementation for the UKB RAP can be found at https://github.com/Cui-yd/SurvSTAARpipeline. The code for generating time-to-event phenotypes in UKB is also available at https://github.com/Cui-yd/ukbbPheSurv.

10 Acknowledgements

This study was supported by National Natural Science Foundation of China (12171318, 12571303), National Key Research and Development Program of China (2023YFC3603200), Shanghai Science and Technology Commission (24JS2810200, 21ZR1436300, 23XD1401900, 23DZ2290600), Medical Engineering Cross Fund of Shanghai Jiao Tong University (YG2023ZD21, YG2023QNA01), Shanghai Rising-Star Program (23YF1421000), Shanghai Municipal Commission of Education (JWAIYB-3), Clinical Research Project of Shanghai Municipal Health Commission in Health Industry (20234Y0285), the Sichuan Science and Technology Program (2024ZYD0102, 2025YFHZ0069), Chengdu Science and Technology Program (2024-YF05-01784-SN).

The simulation computations were conducted on the Siyuan-1 cluster, supported by the Center for High Performance Computing at Shanghai Jiao Tong University. We thank Jianle Sun for his contributions in the early stages of the project and Liu Cheng, Xuyang Ran, and Taeyoung Choi for their valuable discussions on protein structure analysis.

11 Author contributions

Y. C., S. M., Z. L., X. L., and Z. Y. conceived the study. Y. C., S. M., Z. L., X. L., Z. Y., N. Z, and H. C. contributed to methodology development. Y. C., Y. Y., and T. W. performed data analysis. Y. C., S. M., Z. L., and X. L. drafted the manuscript and revised it according to suggestions by the coauthors. All authors critically reviewed the paper, suggested revisions as needed, and approved the final version.

12 Competing interests

The authors declare no competing interests.

References

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., et al. (**2024**). "Accurate structure prediction of biomolecular interactions with AlphaFold 3," Nature, **630**, 493–500. doi:10.1038/s41586-024-07487-w

Andersen, O. M., Monti, G., Jensen, A. M. G., De Waal, M., Hulsman, M., Olsen, J. G., and Holstege, H. (**2023**). "Relying on the relationship with known disease-causing variants in homologous proteins to predict pathogenicity of *SORL1* variants in Alzheimer's disease,." doi:10.1101/2023.02.27.524103

Andersen, O. M., Reiche, J., Schmidt, V., Gotthardt, M., Spoelgen, R., Behlke, J., Von Arnim, C. A. F., et al. (**2005**). "Neuronal sorting protein-related receptor sorLA/LR11 regulates processing of the amyloid precursor protein," Proc. Natl. Acad. Sci., **102**, 13461–13466. doi:10.1073/pnas.0503689102

ARUK Consortium, GERAD/PERADES, CHARGE, ADGC, EADI, Sims, R., Van Der Lee, S. J., Naj, A. C., Bellenguez, C., Badarinarayan, N., et al. (**2017**). "Rare coding variants in *PLCG2*, *ABI3*, and *TREM2* implicate microglial-mediated innate immunity in Alzheimer's disease," Nat. Genet., **49**, 1373–1384. doi:10.1038/ng.3916

Bellenguez, C., Küçükali, F., Jansen, I. E., Kleineidam, L., Moreno-Grau, S., Amin, N., Naj, A. C., et al. (**2022**). "New insights into the genetic etiology of Alzheimer's disease and related dementias," Nat. Genet., **54**, 412–436. doi:10.1038/s41588-022-01024-z

Bi, W., Fritsche, L. G., Mukherjee, B., Kim, S., and Lee, S. (**2020**). "A Fast and Accurate Method for Genome-Wide Time-to-Event Data Analysis and Its Application to UK Biobank," Am. J. Hum. Genet., **107**, 222–233. doi:10.1016/j.ajhg.2020.06.003

Bick, A. G., Metcalf, G. A., Mayo, K. R., Lichtenstein, L., Rura, S., Carroll, R. J., Musick, A., et al. (**2024**). "Genomic data in the all of us research program," Nature, **627**, 340–346. doi:10.1038/s41586-023-06957-x

Breslow, N. E. (**1972**). "Discussion on professor Cox's paper," J. R. Stat. Soc. Ser. B Methodol., **34**, 216–217.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., et al. (**2018**). "The UK Biobank resource with deep phenotyping and genomic data," Nature, **562**, 203–209. doi:10.1038/s41586-018-0579-z

Carss, K., Halldorsson, B. V., Hou, L., Liu, J., and Wheeler, E. (**2025**). "Whole-genome sequencing of 490,640 UK biobank participants," Nature, **645**, 692–701. doi:10.1038/s41586-025-09272-9

Chen, H., Lumley, T., Brody, J., Heard-Costa, N. L., Fox, C. S., Cupples, L. A., and Dupuis, J. (**2014**). "Sequence Kernel Association Test for Survival Traits: SKAT for Survival Traits," Genet. Epidemiol., **38**, 191–197. doi:10.1002/gepi.21791

Cheng, J., Novati, G., Pan, J., Bycroft, C., Žemgulytė, A., Applebaum, T., Pritzel, A., et al. (**2023**). "Accurate proteome-wide missense variant effect prediction with AlphaMissense," Science, **381**, eadg7492. doi:10.1126/science.adg7492

Cox, D. R. (**1972**). “Regression models and life-tables,” J. R. Stat. Soc. Ser. B Methodol., **34**, 187–220.

Davies, R. B. (**1980**). “Algorithm AS 155: The Distribution of a Linear Combination of χ2 Random Variables,” Appl. Stat., **29**, 323. doi:10.2307/2346911

De Deyn, L., and Sleegers, K. (**2025**). “The impact of rare genetic variants on Alzheimer disease,” Nat. Rev. Neurol., **21**, 127–139. doi:10.1038/s41582-025-01062-1

Dey, R., Schmidt, E. M., Abecasis, G. R., and Lee, S. (**2017**). “A Fast and Accurate Algorithm to Test for Binary Phenotypes and Its Application to PheWAS,” Am. J. Hum. Genet., **101**, 37–49. doi:10.1016/j.ajhg.2017.05.014

Dey, R., Zhou, W., Kiiskinen, T., Havulinna, A., Elliott, A., Karjalainen, J., Kurki, M., et al. (**2022**). “Efficient and accurate frailty model approach for genome-wide survival association analysis in large-scale biobanks,” Nat. Commun., **13**, 5437. doi:10.1038/s41467-022-32885-x

Dong, C., Wei, P., Jian, X., Gibbs, R., Boerwinkle, E., Wang, K., and Liu, X. (**2015**). “Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies,” Hum. Mol. Genet., **24**, 2125–2137. doi:10.1093/hmg/ddu733

Fazeli, E., Child, D. D., Bucks, S. A., Stovarsky, M., Edwards, G., Rose, S. E., Yu, C.-E., et al. (**2024**). “A familial missense variant in the Alzheimer’s disease gene SORL1 impairs its maturation and endosomal sorting,” Acta Neuropathol. (Berl.), **147**, 20. doi:10.1007/s00401-023-02670-1

Feuerverger, A. (**1989**). “On the empirical saddlepoint approximation,” Biometrika, **76**, 457–464. doi:10.2307/2336112

Fishilevich, S., Nudel, R., Rappaport, N., Hadar, R., Plaschkes, I., Iny Stein, T., Rosen, N., et al. (**2017**). “GeneHancer: genome-wide integration of enhancers and target genes in GeneCards,” Database, **2017**, bax028. doi:10.1093/database/bax028

Frankish, A., Diekhans, M., Ferreira, A.-M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J. M., et al. (**2019**). “GENCODE reference annotation for the human and mouse genomes,” Nucleic Acids Res., **47**, D766–D773. doi:10.1093/nar/gky955

Halldorsson, B. V., Eggertsson, H. P., Moore, K. H. S., Hauswedell, H., Eiriksson, O., Ulfarsson, M. O., Palsson, G., et al. (**2022**). “The sequences of 150,119 genomes in the UK Biobank,” Nature, **607**, 732–740. doi:10.1038/s41586-022-04965-x

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B. L., et al. (**2012**). “GENCODE: The reference human genome annotation for The ENCODE Project,” Genome Res., **22**, 1760–1774. doi:10.1101/gr.135350.111

He, L., and Kulminski, A. M. (**2020**). “Fast algorithms for conducting large-scale GWAS of age-at-onset traits using Cox mixed-effects models,” Genetics, **215**, 41–58. doi:10.1534/genetics.119.302940

He, L., Loika, Y., Park, Y., Genotype Tissue Expression (GTEx) consortium, Bennett, D. A., Kellis, M., Kulminski, A. M., et al. (**2021**). “Exome-wide age-of-onset analysis reveals exonic variants in *ERN1* and *SPPL2C* associated with Alzheimer’s disease,” Transl. Psychiatry, **11**, 146. doi:10.1038/s41398-021-01263-4

Holstege, H., De Waal, M. W. J., Tesi, N., Van Der Lee, S. J., De Geus, C., Van Spaendonk, R., Vogel, M., et al. (**2023**). "Domain mapping of disease mutations supports genetic testing of specific *SORL1* variants in familial Alzheimer's disease,." doi:10.1101/2023.07.13.23292622

Huang, Y.-F., Gulko, B., and Siepel, A. (**2017**). "Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data," Nat. Genet., **49**, 618–624. doi:10.1038/ng.3810

Jack, C. R., Andrews, J. S., Beach, T. G., Buracchio, T., Dunn, B., Graf, A., Hansson, O., et al. (**2024**). "Revised criteria for diagnosis and staging of Alzheimer's disease: Alzheimer's association workgroup," Alzheimers Dement., **20**, 5143–5169. doi:10.1002/alz.13859

Jansen, I. E., Savage, J. E., Watanabe, K., Bryois, J., Williams, D. M., Steinberg, S., Sealock, J., et al. (**2019**). "Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk," Nat. Genet., **51**, 404–413. doi:10.1038/s41588-018-0311-9

Kinsella, R. J., Kahari, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., Almeida-King, J., et al. (**2011**). "Ensembl BioMarts: a hub for data retrieval across taxonomic space," Database, **2011**, bar030. doi:10.1093/database/bar030

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., and Shendure, J. (**2014**). "A general framework for estimating the relative pathogenicity of human genetic variants," Nat. Genet., **46**, 310–315. doi:10.1038/ng.2892

Lee, S., Abecasis, G. R., Boehnke, M., and Lin, X. (**2014**). "Rare-variant association analysis: study designs and statistical tests," Am. J. Hum. Genet., **95**, 5–23. doi:10.1016/j.ajhg.2014.06.009

Li, B., and Leal, S. M. (**2008**). "Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data," Am. J. Hum. Genet., **83**, 311–321. doi:10.1016/j.ajhg.2008.06.024

Li, X., Li, Z., Zhou, H., Gaynor, S. M., Liu, Y., Chen, H., Sun, R., et al. (**2020**). "Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale," Nat. Genet., **52**, 969–983. doi:10.1038/s41588-020-0676-4

Li, X., Quick, C., Zhou, H., Gaynor, S. M., Liu, Y., Chen, H., Selvaraj, M. S., et al. (**2023**). "Powerful, scalable and resource-efficient meta-analysis of rare variant associations in large whole genome sequencing studies," Nat. Genet., **55**, 154–164. doi:10.1038/s41588-022-01225-6

Li, X., Wood, A. R., Yuan, Y., Zhang, M., Huang, Y., Hawkes, G., Beaumont, R. N., et al. (**2025**). "Streamlining large-scale genomic data management: Insights from the UK Biobank whole-genome sequencing data," Cell Genomics, **1**, 101009. doi:10.1016/j.xgen.2025.101009

Li, Z., Li, X., Liu, Y., Shen, J., Chen, H., Zhou, H., Morrison, A. C., et al. (**2019**). "Dynamic scan procedure for detecting rare-variant association regions in whole-genome sequencing studies," Am. J. Hum. Genet., **104**, 802–814. doi:10.1016/j.ajhg.2019.03.002

Li, Z., Li, X., Zhou, H., Gaynor, S. M., Selvaraj, M. S., Arapoglou, T., Quick, C., et al. (**2022**). "A framework for detecting noncoding rare-variant associations of large-scale whole-genome sequencing studies," Nat. Methods, **19**, 1599–1611. doi:10.1038/s41592-022-01640-x

Lin, D. Y. (**2007**). "On the Breslow estimator," Lifetime Data Anal., **13**, 471–480. doi:10.1007/s10985-007-9048-y

Liu, H., Tang, Y., and Zhang, H. H. (**2009**). "A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables," Comput. Stat. Data Anal., **53**, 853–856. doi:10.1016/j.csda.2008.11.025

Liu, Y., Chen, S., Li, Z., Morrison, A. C., Boerwinkle, E., and Lin, X. (**2019**). "ACAT: A Fast and Powerful p Value Combination Method for Rare-Variant Analysis in Sequencing Studies," Am. J. Hum. Genet., **104**, 410–421. doi:10.1016/j.ajhg.2019.01.002

Madsen, B. E., and Browning, S. R. (**2009**). "A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic," PLoS Genet., **5**, e1000384. doi:10.1371/journal.pgen.1000384

Mbatchou, J., Barnard, L., Backman, J., Marcketta, A., Kosmicki, J. A., Ziyatdinov, A., Benner, C., et al. (**2021**). "Computationally efficient whole-genome regression for quantitative and binary traits," Nat. Genet., **53**, 1097–1103. doi:10.1038/s41588-021-00870-7

Mehmedbasic, A., Christensen, S. K., Nilsson, J., Rüetschi, U., Gustafsen, C., Poulsen, A. S. A., Rasmussen, R. W., et al. (**2015**). "SorLA Complement-type Repeat Domains Protect the Amyloid Precursor Protein against Processing," J. Biol. Chem., **290**, 3359–3376. doi:10.1074/jbc.M114.619940

Moreno-Grau, S., De Rojas, I., Hernández, I., Quintela, I., Montrreal, L., Alegret, M., Hernández-Olasagarre, B., et al. (**2019**). "Genome-wide association analysis of dementia and its clinical endophenotypes reveal novel loci associated with Alzheimer's disease and three causality networks: The GR@ACE project," Alzheimers Dement., **15**, 1333–1347. doi:10.1016/j.jalz.2019.06.4950

Morgenthaler, S., and Thilly, W. G. (**2007**). "A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test (CAST)," Mutat. Res. Mol. Mech. Mutagen., **615**, 28–56. doi:10.1016/j.mrfmmm.2006.09.003

Morris, A. P., and Zeggini, E. (**2010**). "An evaluation of statistical approaches to rare variant analysis in genetic association studies," Genet. Epidemiol., **34**, 188–193. doi:10.1002/gepi.20450

Nichols, E., Steinmetz, J. D., Vollset, S. E., Fukutaki, K., Chalek, J., Abd-Allah, F., Abdoli, A., et al. (**2022**). "Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: An analysis for the global burden of disease study 2019," Lancet Public Health, **7**, e105–e125. doi:10.1016/S2468-2667(21)00249-8

Ojavee, S. E., Kousathanas, A., Trejo Banos, D., Orliac, E. J., Patxot, M., Läll, K., Mägi, R., et al. (**2021**). "Genomic architecture and prediction of censored time-to-event phenotypes with a Bayesian genome-wide analysis," Nat. Commun., **12**, 2337. doi:10.1038/s41467-021-22538-w

Pedersen, E. M., Agerbo, E., Plana-Ripoll, O., Steinbach, J., Krebs, M. D., Hougaard, D. M., Werge, T., et al. (**2023**). "ADuLT: An efficient and robust time-to-event GWAS," Nat. Commun., **14**, 5553. doi:10.1038/s41467-023-41210-z

Povysil, G., Petrovski, S., Hostyk, J., Aggarwal, V., Allen, A. S., and Goldstein, D. B. (**2019**). "Rare-variant collapsing analyses for complex traits: guidelines and applications," Nat. Rev. Genet., **20**, 747–759. doi:10.1038/s41576-019-0177-4

Rizvi, A. A., Karaesmen, E., Morgan, M., Preus, L., Wang, J., Sovic, M., Hahn, T., et al. (**2019**). "gwasurvivr: an R package for genome-wide survival analysis," (R. Schwartz, Ed.) Bioinformatics, **35**, 1968–1970. doi:10.1093/bioinformatics/bty920

Rogers, M. F., Shihab, H. A., Mort, M., Cooper, D. N., Gaunt, T. R., and Campbell, C. (**2018**). "FATHMM-XF: accurate prediction of pathogenic point mutations via extended features," (J. Hancock, Ed.) Bioinformatics, **34**, 511–513. doi:10.1093/bioinformatics/btx536

Schaffner, S. F., Foo, C., Gabriel, S., Reich, D., Daly, M. J., and Altshuler, D. (**2005**). "Calibrating a coalescent simulation of human genome sequence variation," Genome Res., **15**, 1576–1583. doi:10.1101/gr.3709305

Schwartzentruber, J., Cooper, S., Liu, J. Z., Barrio-Hernandez, I., Bello, E., Kumasaka, N., Young, A. M. H., et al. (**2021**). "Genome-wide meta-analysis, fine-mapping and integrative prioritization implicate new Alzheimer's disease risk genes," Nat. Genet., **53**, 392–402. doi:10.1038/s41588-020-00776-w

Sollis, E., Mosaku, A., Abid, A., Buniello, A., Cerezo, M., Gil, L., Groza, T., et al. (**2023**). "The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource," Nucleic Acids Res., **51**, D977–D985. doi:10.1093/nar/gkac1010

Spence, J. P., Mostafavi, H., Ota, M., Milind, N., Gjorgjieva, T., Smith, C. J., Simons, Y. B., et al. (**2025**). "Specificity, length and luck drive gene rankings in association studies," Nature, **0**, 1–8. doi:10.1038/s41586-025-09703-7

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., et al. (**2015**). "UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age," PLoS Med., **12**, e1001779. doi:10.1371/journal.pmed.1001779

Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres, R., Taliun, S. A. G., et al. (**2021**). "Sequencing of 53,831 diverse genomes from the NHLBI TOPMed program," Nature, **590**, 290–299. doi:10.1038/s41586-021-03205-y

Terry M. Therneau (**2024**). "A Package for Survival Analysis in R.," Retrieved from https://CRAN.R-project.org/package=survival

Terry M. Therneau and Patricia M. Grambsch (**2000**). *Modeling Survival Data: Extending the Cox Model*, Springer, New York.

Teufel, F., Almagro Armenteros, J. J., Johansen, A. R., Gíslason, M. H., Pihl, S. I., Tsirigos, K. D., Winther, O., et al. (**2022**). "SignalP 6.0 predicts all five types of signal peptides using protein language models," Nat. Biotechnol., **40**, 1023–1025. doi:10.1038/s41587-021-01156-3

The All of Us Research Program Investigators (**2019**). "The 'all of us' research program," N. Engl. J. Med., **381**, 668–676. doi:10.1056/NEJMsr1809937

The ENCODE Project Consortium (**2012**). "An integrated encyclopedia of DNA elements in the human genome," Nature, **489**, 57–74. doi:10.1038/nature11247

The ENCODE Project Consortium, Abascal, F., Acosta, R., Addleman, N. J., Adrian, J., Afzal, V., Ai, R., et al. (**2020**). "Expanded encyclopaedias of DNA elements in the human and mouse genomes," Nature, **583**, 699–710. doi:10.1038/s41586-020-2493-4

The FANTOM Consortium and the RIKEN PMI and CLST (DGT) (**2014**). "A promoter-level mammalian expression atlas," Nature, **507**, 462–470. doi:10.1038/nature13182

The FANTOM Consortium, Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., et al. (**2014**). "An atlas of active enhancers across human cell types and tissues," Nature, **507**, 455–461. doi:10.1038/nature12787

Wainschtein, P., Zhang, Y., Schwartzentruber, J., Kassam, I., Sidorenko, J., Fiziev, P. P., Wang, H., et al. (**2025**). "Estimation and mapping of the missing heritability of human phenotypes," Nature, **0**, 1–9. doi:10.1038/s41586-025-09720-6

Wightman, D. P., Jansen, I. E., Savage, J. E., Shadrin, A. A., Bahrami, S., Holland, D., Rongve, A., et al. (**2021**). "A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease," Nat. Genet., **53**, 1276–1282. doi:10.1038/s41588-021-00921-z

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (**2011**). "Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test," Am. J. Hum. Genet., **89**, 82–93. doi:10.1016/j.ajhg.2011.05.029

Xu, H., Ma, Y., Xu, L., Li, Y., Liu, Y., Li, Y., Zhou, X., et al. (**2025**). "SPAGRM: effectively controlling for sample relatedness in large-scale genome-wide association studies of longitudinal traits," Nat. Commun., **16**, 1413. doi:10.1038/s41467-025-56669-1

Zhang, Y., Chen, H., Li, R., Sterling, K., and Song, W. (**2023**). "Amyloid β-based therapy for Alzheimer's disease: challenges, successes and future," Signal Transduct. Target. Ther., **8**, 248. doi:10.1038/s41392-023-01484-7

Zhou, H., Arapoglou, T., Li, X., Li, Z., and Lin, X. (**2022**). "FAVOR essential database.," doi:10.7910/DVN/1VGTJI

Zhou, H., Arapoglou, T., Li, X., Li, Z., Zheng, X., Moore, J., Asok, A., et al. (**2023**). "FAVOR: functional annotation of variants online resource and annotator for variation across the human genome," Nucleic Acids Res., **51**, D1300–D1311. doi:10.1093/nar/gkac966

Zhou, W., Nielsen, J. B., Fritsche, L. G., Dey, R., Gabrielsen, M. E., Wolford, B. N., LeFaive, J., et al. (**2018**). "Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies," Nat. Genet., **50**, 1335–1341. doi:10.1038/s41588-018-0184-y

Table 1. Rare variant analysis results for gene-centric coding regions in Alzheimer's disease (UK Biobank 500k WGS data, n=458,773).

Gene	Chr	Category	#SNV	cMAC	SurvSTAAR-O (Unconditional)	SurvSTAAR-O (Conditional)	Variants adjusted
TREM2	6	Missense	237	15886	5.26×10^{-13}	2.26×10^{-1}	rs75932628, rs143332484
SORL1	11	Disruptive missense	517	5240	1.73×10^{-7}	1.80×10^{-7}	rs11218343, rs74685827, rs117807585, rs4559697, rs509334
SORL1	11	Putative loss-of-function and disruptive missense	590	5705	2.04×10^{-9}	1.97×10^{-9}	rs11218343, rs74685827, rs117807585, rs4559697, rs509334
BCAM	19	Synonymous	274	13650	3.12×10^{-7}	9.81×10^{-1}	rs429358, rs7412
TOMM40	19	Missense	258	3495	7.70×10^{-13}	1.77×10^{-3}	rs429358, rs7412
APOE	19	Missense	297	6544.005	1.78×10^{-8}	4.45×10^{-3}	rs429358, rs7412
CLPTM1	19	Missense	522	9310	2.15×10^{-10}	4.68×10^{-3}	rs429358, rs7412

NOTE: This table presents results for the unconditionally significant genes (SurvSTAAR-O *p*-value $< 5.00\times10^{-7}$), determined using Bonferroni correction for multiple comparisons (0.05/[20,000×5]) across five different coding masks in protein-coding genes. Chr: chromosome; Category: functional category; #SNV: number of rare variants (MAF < 1%) of the particular functional category in the gene; cMAC: cumulative minor allele count; SurvSTAAR-O: *p*-value of SurvSTAAR-O test.

Table 2. Rare variant analysis results for gene-centric noncoding regions in Alzheimer's disease (UK Biobank 500k WGS data, n=458,773).

Gene	Chr	Category	#SNV	cMAC	SurvSTAAR-O (Unconditional)	SurvSTAAR-O (Conditional)	Variants (adjusted)
TREM2	6	promoter_CAGE	150	13393	1.82×10^{-13}	2.40×10^{-1}	rs75932628, rs143332484
TREM2	6	promoter_DHS	610	26649	2.29×10^{-12}	8.98×10^{-1}	rs75932628, rs143332484
TREM2	6	enhancer_CAGE	125	12427	6.03×10^{-13}	7.52×10^{-1}	rs75932628, rs143332484
TREM2	6	enhancer_DHS	723	26567	1.18×10^{-12}	7.10×10^{-1}	rs75932628, rs143332484
BCAM	19	UTR	535	16230.01	1.49×10^{-9}	7.04×10^{-4}	rs429358, rs7412
BCAM	19	enhancer_DHS	1055	30989	1.24×10^{-8}	2.09×10^{-1}	rs429358, rs7412
NECTIN2	19	enhancer_CAGE	1050	41914	1.14×10^{-9}	3.25×10^{-3}	rs429358, rs7412
NECTIN2	19	enhancer_DHS	4130	144067	1.43×10^{-8}	3.61×10^{-1}	rs429358, rs7412
APOE	19	promoter_CAGE	505	9127	3.77×10^{-8}	1.09×10^{-1}	rs429358, rs7412
APOE	19	promoter_DHS	793	16055	3.49×10^{-8}	3.72×10^{-1}	rs429358, rs7412
APOC1	19	promoter_DHS	1243	37855	1.44×10^{-10}	1.38×10^{-1}	rs429358, rs7412
APOC2	19	enhancer_DHS	1765	44129	5.22×10^{-10}	8.42×10^{-1}	rs429358, rs7412
CLPTM1	19	promoter_DHS	1046	23374	1.14×10^{-8}	7.21×10^{-1}	rs429358, rs7412

NOTE: This table presents results for the unconditionally significant genes (SurvSTAAR-O *p*-value $< 3.57\times10^{-7}$, determined using Bonferroni correction for multiple comparisons (0.05/[20,000×7]) across seven different noncoding masks in protein-coding genes. Chr: chromosome; Category: functional category; #SNV: number of rare variants (MAF < 1%) of the particular functional category in the gene; cMAC: cumulative minor allele count; SurvSTAAR-O: *p*-value of SurvSTAAR-O test.

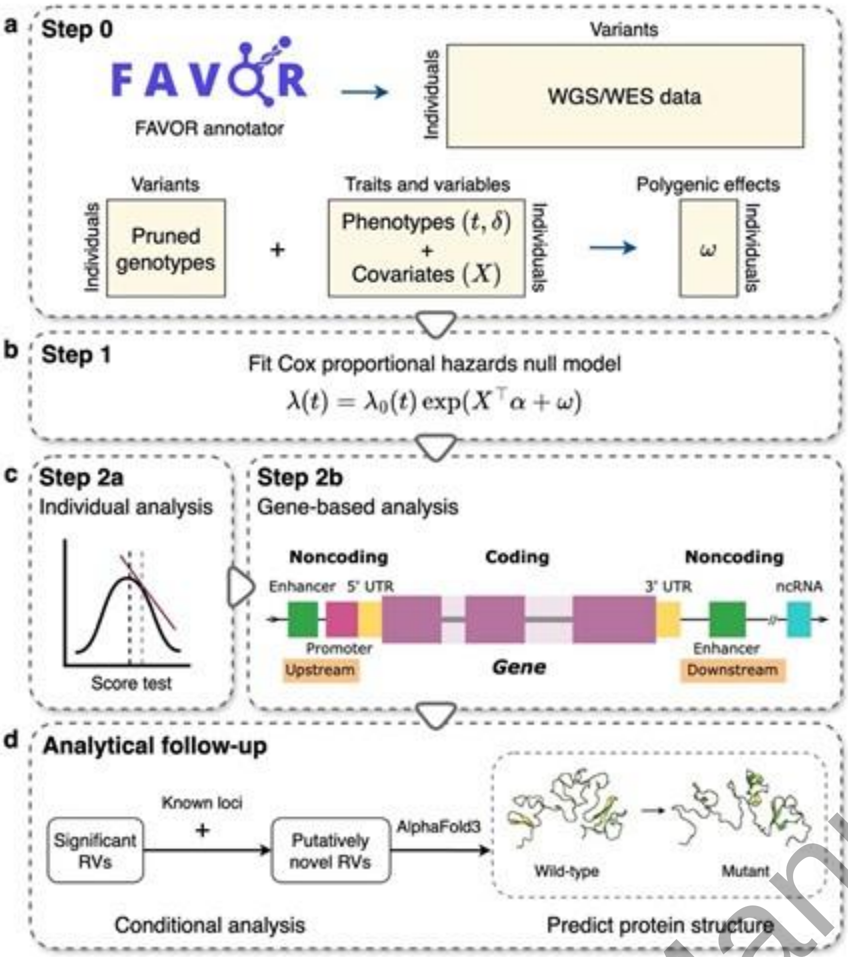

Figure 1. Workflow of SurvSTAAR. (a) Step 0 of SruvSTAAR framework, annotate all variants through FAVORannotator, and calculate polygenic scores. **(b)** Step 1 of SruvSTAAR framework, fit Cox proportional hazards null model. **(c)** Step 2 of SruvSTAAR framework, step 2a: individual analysis, calculate score test statistics for each genetic variants; step 2b: set-based analysis for rare variants (including gene-centric coding, noncoding, and ncRNA rare variants). **(d)** Analytical follow-up. Perform conditional analysis on genome-wide significant RV sets. For putatively novel RV sets, use AlphaFold3 to predict mutant protein structures.

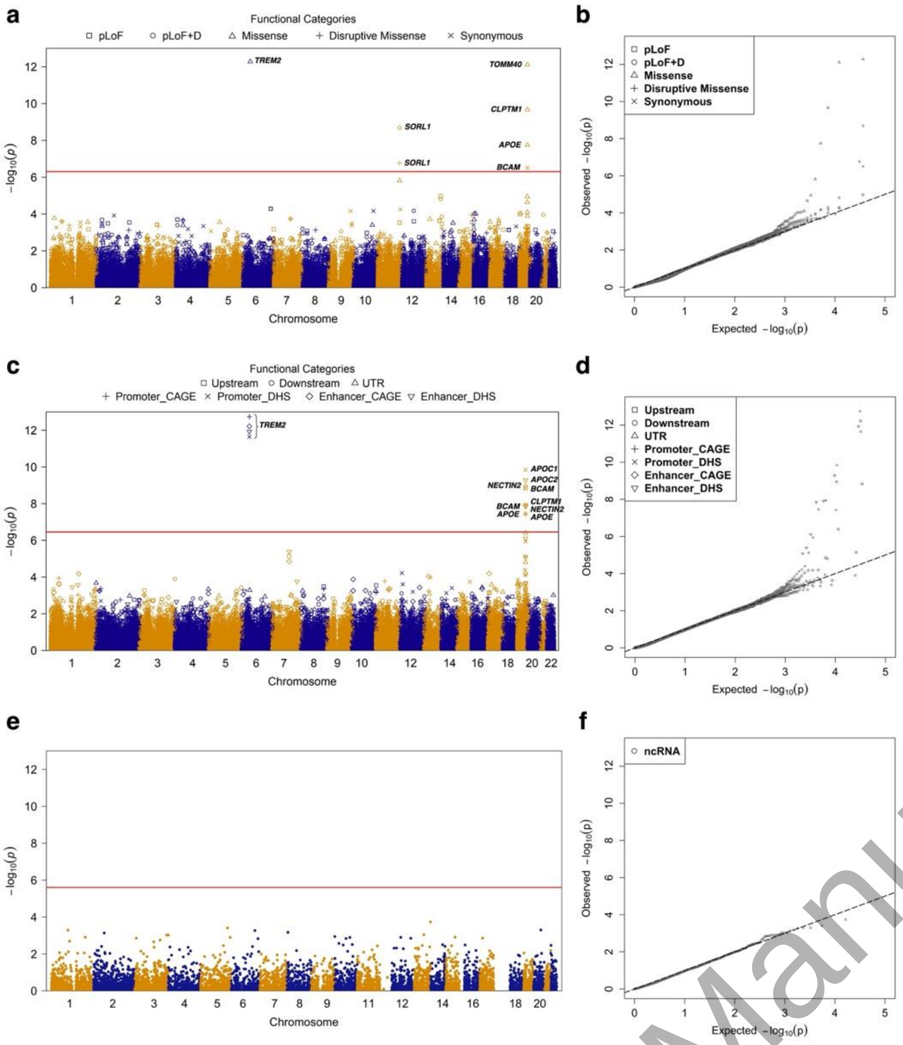

Figure 2. Manhattan plots and Q-Q plots for unconditional gene-centric coding, noncoding, and ncRNA rare variants analysis for Alzheimer's disease using UK Biobank 500k WGS data (n=458,773). (a) Manhattan plots for five different unconditional gene-centric coding masks across protein-coding genes. **(b)** Quantile-quantile plot for five different unconditional gene-centric coding masks across protein-coding genes. **(c)** Manhattan plots for seven different unconditional gene-centric noncoding masks across protein-coding genes. **(d)** Quantile-quantile plot for seven different unconditional gene-centric noncoding masks across protein-coding genes. **(e)** Manhattan plots for unconditional gene-centric noncoding masks across ncRNA genes. **(f)** Quantile-quantile plot for unconditional gene-centric noncoding masks across ncRNA genes.

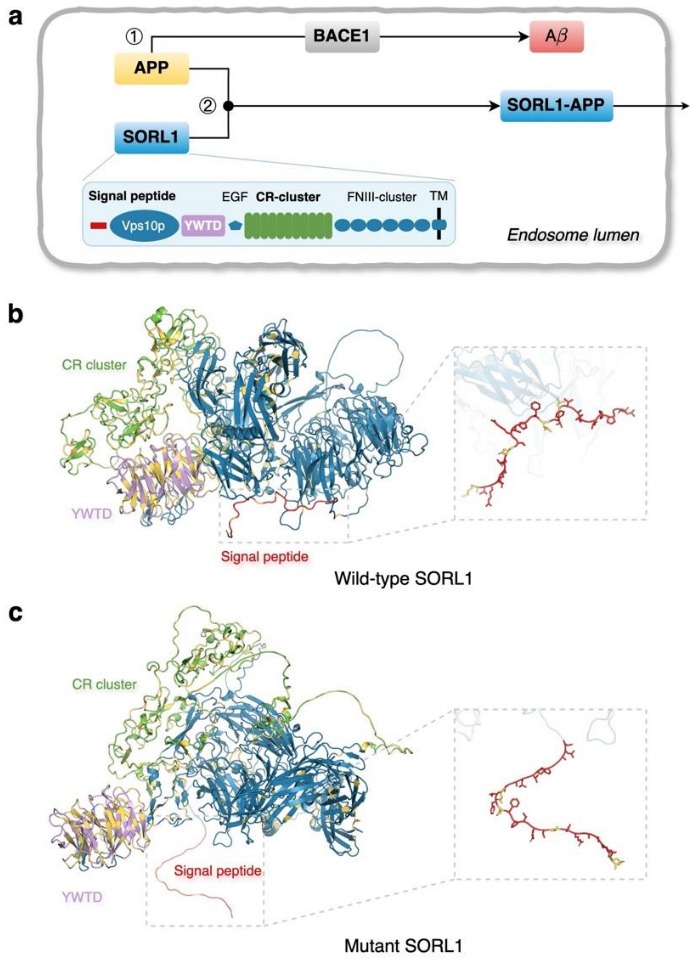

Figure 3. The role of SORL1 and structural changes caused by missense mutations. (a) The structure of SORL1 and its role in APP processing. The structural elements include the signal peptide, vacuolar protein sorting 10 protein (Vps10p) domain, YWTD β-propeller domain, epidermal growth factor repeat (EGF), complement-type repeats (CR) cluster, fibronectin type III (FNIII) cluster, and transmembrane. *Pathway 1*: APP in late endocytic compartments is cleaved by beta site APP cleaving enzyme 1 (BACE1) into Aβ. *Pathway 2*: SORL1 acts as a sorting receptor that binds APP in the endosome, redirecting it into the Golgi. **(b)** The wild-type SORL1 structure. Protein domains are colored the same as in **(a)**, with mutated amino acids highlighted in yellow. **(c)** The mutated SORL1 structure. Protein domains are colored the same as in (**a**), with mutated amino acids highlighted in yellow.