Testing for association with rare variants in the coding and non-coding genome: RAVA-FIRST, a new approach based on CADD deleteriousness score

Ozvan Bocher[1,*], Thomas E. Ludwig[1,2], Gaëlle Marenne[1], Jean-François Deleuze[3], Suryakant Suryakant[4], Jacob Odeberg[5,6], Pierre-Emmanuel Morange[7], David-Alexandre Trégouët[4], Hervé Perdry[8], Emmanuelle Génin[1,2]

[1]Univ Brest, Inserm, EFS, UMR 1078, GGB, F-29200 Brest, France

[2]CHU Brest, Brest, France

[3]Centre National de Recherche en Génomique Humaine CNRGH, Institut de Biologie François Jacob, Université Paris Saclay, CEA, 2 rue Gaston Crémieux CP 5721, 91057 Evry, France

[4]University of Bordeaux, Inserm, Bordeaux Population Health Research Center, team ELEANOR, UMR 1219, 33000 Bordeaux, France

[5]Science for Life Laboratory, Department of Protein Science, CBH, KTH Royal Institute of Technology, Stockholm, Sweden

[6]Department of Clinical Medicine, Faculty of Health Science, The Arctic University of Tromsö, Tromsö, Norway

[7]Aix Marseille Univ, INSERM, INRAE, C2VN, Marseille, France

[8]CESP Inserm, U1018, UFR Médecine, Univ Paris-Sud, Université Paris-Saclay, Villejuif, France

*Corresponding author

Email: bocherozvan@gmail.com (OB)

Abstract

Rare variant association tests (RVAT) have been developed to study the contribution of rare variants widely accessible through high-throughput sequencing technologies. RVAT require to aggregate rare variants in testing units and to filter variants to retain only the most likely causal ones. In the exome, genes are natural testing units and variants are usually filtered based on their functional consequences. However, when dealing with whole-genome sequence (WGS) data, both steps are challenging. No natural biological unit is available for aggregating rare variants. Sliding windows procedures have been proposed to circumvent this difficulty, however they are blind to biological information and result in a large number of tests.

We propose a new strategy to perform RVAT on WGS data: “RAVA-FIRST” (RAre Variant Association using Functionally-InfoRmed STeps) comprising three steps. (1) New testing units are defined genome-wide based on functionally-adjusted Combined Annotation Dependent Depletion (CADD) scores of variants observed in the GnomAD populations, which are referred to as “CADD regions”. (2) A region-dependent filtering of rare variants is applied in each CADD region. (3) A functionally-informed burden test is performed with sub-scores computed for each genomic category within each CADD region. Both on simulations and real data, RAVA-FIRST was found to outperform other WGS-based RVAT. Applied to a WGS dataset of venous thromboembolism patients, we identified an intergenic region on chromosome 18 that is enriched for rare variants in early-onset patients and that was that was missed by standard sliding windows procedures.

RAVA-FIRST enables new investigations of rare non-coding variants in complex diseases, facilitated by its implementation in the R package Ravages.

Author Summary

Technological progresses have made possible whole genome sequencing at an unprecedented scale, opening up the possibility to explore the role of genetic variants of low frequency in common diseases. The challenge is now methodological and requires the development of novel methods and strategies to analyse sequencing data that are not limited to assessing the role of coding variants. With RAVA-FIRST, we propose a novel strategy to investigate the role of rare variants in the whole-genome that takes benefit from biological information. Especially, RAVA-FIRST relies on testing units that go beyond genes to gather rare variants in the association tests. In this work, we show that this new strategy presents several advantages compared to existing methods. RAVA-FIRST offers an easy and straightforward analysis of genome-wide rare variants, especially the intergenic ones which are frequently left behind, making it a promising tool to get a better understanding of the biology of complex diseases.

Introduction

With advance in sequencing technologies, it is now possible to explore the role of rare genetic variants in complex diseases. Different rare variant association tests (RVAT) have been developed that gather rare variants into testing units and compare rare variant content in these testing units between cases and controls (1–3). While the impact of rare variants has already been shown in several complex diseases (4–6), RVAT face two key challenges: (i) the definition of the testing units and (ii) the selection of the qualifying rare variants to include in these units. The proportion of causal variants in the testing units being a major driver of power, especially for burden tests, it is indeed important to ensure that qualifying variants are enriched in variants likely to have some functional impact (3,7). When exome analyses are undertaken, rare variants are most often grouped by genes and included in the analysis depending on their impact on the corresponding protein (8,9). Nevertheless, the gene definition is not always optimal as differences in rare variants burden between cases and controls could sometimes only be found in a sub-region of a gene. This is for example the case in the *RNF213* gene where an enrichment in rare variants located in the C-terminal region is found in Moyamoya cases (10). Defining testing units and qualifying variants is also much more challenging in the non-coding genome due to the lack of defined genomic elements and the higher difficulty to predict the functional impact of non-coding variants (11). It is yet a question of interest as several studies have shown the importance of rare non-coding variants in the development of complex diseases (12–14). Functional elements such as enhancers or promoters can be used as testing units (5,15,16) but they prevent the analysis of all rare variants in the genome and can be too small to get a sufficient number of rare variants for association analysis. On the other hand, sliding windows procedures such as SCAN-G (17) or WGSCAN (18) can be used to test for association over the whole genome. Nevertheless, they present several limits including the window definition that is arbitrary and blind to biological information, the high number of tests and the associated computation time. With overlapping windows, there is also a strong correlation between tests performed in the different testing units that requires the use of

82 permutation procedures to account for multiple testing. Finally, to filter rare variants in the testing
83 units, pathogenicity scores are often used but without guidelines on which score to use and which
84 threshold to apply.

85 In this paper, we propose RAVA-FIRST (RAre Variant Association using Functionally InfoRmed STeps), a
86 new strategy for analysing rare variants in the coding and the non-coding genome that addresses the
87 previous issues. First, we provide pre-defined testing units in the whole genome called “CADD regions”
88 based on the Combined Annotation Dependent Depletion (CADD) scores of deleteriousness of variants
89 observed in the GnomAD general population. These regions prevent the use of sliding windows
90 procedures while enabling the study of rare variants in the whole genome. Second, we propose a
91 filtering approach based on CADD scores with region-dependant thresholds to represent the genetic
92 context of each CADD region and avoid the use of a fix threshold along the genome. Finally, we
93 integrate functional information into the burden test to detect an accumulation of rare variants in
94 specific genomic categories within CADD regions. Through a statistical description of these testing
95 units, we show that they preserve the integrity of the majority of functional elements in the genome.
96 We also show that the RAVA-FIRST filtering strategy enables a better discrimination between
97 functional and non-functional variants within the testing units. We applied RAVA-FIRST to real whole-
98 genome sequencing data from individuals with venous thromboembolism (VTE) and detected an
99 intergenic association signal that would have been missed with sliding windows and a classical filtering
100 of rare variants. RAVA-FIRST is implemented in the R package Ravages available on the CRAN and
101 maintained on github (19,20).

102

103

Description of the Method

RAVA-FIRST is developed to test for association with rare variants in the whole genome. It deals with all steps from the definition of testing units and the filtering of rare variants, to the association test accounting for functional information. The main steps are represented in S1 Fig and further details are presented hereafter.

Testing units in RAVA-FIRST: the CADD regions

Following Havrilla et al. (2019) (21), we seek to identify some genomic regions that were significantly depleted in functional variants to use them as testing units in RVAT. For that purpose, Havrilla et al. (2019) defined "constrained coding regions" (CCR) as exonic regions where no important functional variation (defined as being at least missense) was found in the general population of GnomAD (22). In our experience, two limits prevent the direct use of CCR as testing units in the whole genome: they are too small to gather a sufficient number of rare variants (224 bp being the maximum length of a CCR) and their definition relies on the consequence of the variants on the translated protein, not available in the non-coding genome. We therefore decided to expand the proposed approach by estimating the functionality of variants through CADD scores (23). CADD scores were chosen because of their availability for every substitution in the genome and because they rank well in the comparison test of functional annotation tools (24).

Coding variants tend to present higher CADD values than non-coding variants (23). A selection based on a CADD threshold would therefore result in a majority of coding variants selected. In order to avoid this pattern, we adjusted the RAW CADD scores on a PHRED scale within each of three genomic categories: "coding", "regulatory" and "intergenic" regions. Coding regions correspond to CCDS (25) and represent 1.2% of the genome. Regulatory regions represent 44.3% of the genome and are defined by the union of introns, 5' and 3' UTR, promoters and enhancers, all being involved in gene regulation (26). Enhancers and promoters have been obtained with the SCREEN tool from ENCODE which enables

the definition of a large number of regulatory elements in diverse cell types (27). Finally, intergenic regions correspond to all regions not being described as coding or regulatory regions, representing 54.5% of the genome. More details are given in the Supporting Information.

Adjusted CADD scores were used to select the variants that will bound the "CADD regions". First, we selected the variants with an adjusted CADD score greater than 20, that is the top 1% of variants with the highest predicted functional impact within each of the three genomic categories. Then, among those variants, only the ones observed at least two times in GnomAD r2.0.1 genomes were further selected and used as boundaries of CADD regions. For CADD regions to be used as testing units in RVAT, they need to be large enough to contain several rare variants. Contiguous small regions of less than 10 kb were therefore grouped together to form clusters of variants with high adjusted CADD scores. Non-sequenced regions and low-covered regions in GnomAD containing potential important functional variants were excluded from CADD regions, leading to gaps within CADD regions of at least one base pair (i.e. no CADD region overlap them to avoid artificially long regions due to a lack of variants in GnomAD). Finally, CADD regions are only defined for regions where CADD scores are available (removing among others centromeres and telomeres). Note that CADD regions can overlap different genomic categories (coding, regulatory or intergenic). More details about the steps and parameters used for the definition of CADD regions are presented in the Supporting Information.

The RAVA-FIRST filtering strategy

In addition to the definition of new testing units in the whole genome, we propose a new filtering strategy in RAVA-FIRST to select qualifying variants. Using gene-specific CADD thresholds rather than a fixed threshold for all genes was previously found to improve prediction (28). Building on the same idea, we defined thresholds that are specific to each CADD region. To define these region-specific thresholds, we derived the median of all adjusted CADD scores of variants observed at least two times in GnomAD in each CADD region. This value represents the median score level that is tolerated in the

general population within each CADD region. Adjusted CADD scores refer here to the PHRED CADD scores computed respectively for coding, regulatory and intergenic genomic categories as defined before. Qualifying variants are then defined as rare variants with an adjusted CADD score above the threshold specific to their region. Note that because CADD scores are only available for SNVs, other types of variants are excluded from the analyses.

Burden test in RAVA-FIRST: taking into account functional information

As mentioned before, several of the CADD regions overlap different genomic categories (coding, regulatory or intergenic, Figs S1 and S3). As the effects of variants belonging to these different genomic categories may not be the same, we extended the burden test defined as:

$$\ln\frac{P(Y_j=1)}{P(Y_j=0)} = \beta_0 + \beta_{Cov}X_{Cov} + \beta_G X_G$$

With Y_j the vector of phenotypes for the n individuals: 0 for the group of controls and 1 for the group of cases. β_0 represents the intercept of the model and X_{Cov} the matrix of covariates (if any) with their associated effect, β_{Cov}. β_G corresponds to the estimated effect of the burden X_G, computed for example using WSS (1) which corresponds to a weighted sum of rare alleles based on their frequency, the rarest alleles having the highest weights.

To take into account functional information, we integrated a sub-score for each genomic category into the regression model, similarly to the analysis of rare and frequent variants proposed by Li and Leal (2008) (7):

$$\ln\frac{P(Y_j=1)}{P(Y_j=0)} = \beta_0 + \beta_{Cov}X_{Cov} + \sum_{G=\{cod;reg;inter\}} \beta_G X_G$$

Sub-scores X_G are constructed for each genomic category within a CADD region, with at most three sub-scores (coding, regulatory or intergenic). The p-value can be determined using a likelihood ratio

test comparing this model to the null model where the sub-scores are not included. This sub-score analysis, also called RAVA-FIRST burden test, is also available for continuous and for categorical phenotypes using the extension of burden tests developed in Bocher et al. (2019) (19). The RAVA-FIRST burden test coupled with the region-specific filtering on the adjusted CADD score enables to keep the most important functional variants within each genomic category and to take into account those categories in the association test while performing only one test by CADD region.

Verification and Comparison

Statistics on CADD regions and comparison with genomic elements

A total of 135,224 CADD regions were defined covering 93.2% of the genome (in build GRCh37). Among CADD regions, several are very small in size, despite our approach to combine small regions, due to the removal of low-covered regions, preventing their use in RVAT. We therefore decided to focus on the 106,251 CADD regions larger than 1kb, which cover 93% of the genome. Among those CADD regions, 28.3% span only one type of genomic category, 58.5% span two of the three types of genomic categories, and 13.2% overlap the three genomic categories (S3 Fig). Some CADD regions are extremely large, mainly around the centromeres (Table 1). About 80% of CADD regions have a size between 5 and 50 kb with a mean of 25 kb, making them completely compatible with the size of genes commonly used as testing units used in RVAT.

Table 1: Summary statistics of the lengths of CADD regions (larger than 1 kb)

	Quantiles					Mean
	0%	25%	50%	75%	100%	
Length (kb)	1	10.790	16.579	29.116	1,731.228	25.224

We then compared the position of genomic elements relative to the defined CADD regions (Table in S1 Table shows how the different genomic elements have been obtained). A large majority of genomic elements are entirely included into a single CADD region and thus their integrity is preserved (Table 2). This is expected as all these genomic elements are substantially smaller than the CADD regions and therefore have a high probability of being included in a CADD region. For larger elements such as introns or lncRNA, the percentage decreases but remains high (more than 80% of lncRNA are overlapped by at most 2 CADD regions). The genomic elements spanning more than one CADD region are on average longer than the ones being entirely included into a single CADD region. However, when comparing CCR and CADD regions, it is interesting to note that the CCRs entirely encompassed within a single CADD region are the longest ones that should also represent the most constrained regions.

Table 2: Percentage of genomic elements entirely encompassed within a CADD region

Exon CCDS	Protein domains	CCR	Introns/UTR	Enh-Prom		Silencers	CTCF	lncRNA
				DECRES	ENCODE			
97.8%	81.8%	99.2%	85.9%	93.1%	96.4%	95.1%	95.8%	65.5%

Performance of RAVA-FIRST filtering based on adjusted CADD scores

To assess the performance of the adjusted CADD scores and the RAVA-FIRST filtering, we evaluated its capacity in discriminating benign from pathogenic variants using the Clinvar database (29). We computed true positive rate (TPR), true negative rate (TNR) and precision for the RAVA-FIRST filtering and compared the results to the ones obtained by applying a fixed CADD threshold of 10, 15 or 20 on variants annotated with CADD scores v1.4. After the selection of rare variants included in RVAT (see the Supporting Information), the dataset of analysis contains 70,931 variants of which 25,931 are benign and 45,000 are pathogenic. All filtering strategies show a very high TPR (Fig 1A), meaning that the majority of pathogenic variants would be selected as qualifying variants for RVAT. The TNR

increases with the increasing CADD score threshold which is expected as less variants, and therefore less benign variants, are included in the analysis. The RAVA-FIRST filtering shows the highest TNR and the highest precision. While the TPR value is extremely important to select the most probable causal variants in RVAT, it is also important to have a high TNR value, otherwise the signal will be diluted by a high proportion of non-causal variants. The precision value summarises the TPR and TNR parameters and therefore, to a certain extent, is representative of the percentage of causal variants among selected variants. Therefore, we show that the RAVA-FIRST filtering strategy is the most accurate to select qualifying rare variants for RVAT. Focusing on the coding genome, we also compared the performance of RAVA-FIRST filtering approach against two others approaches classically used on genes as testing units: (1) filter for variants with a functional impact expected to change the protein (“missense_variant", "missense_variant&splice_region_variant", "splice_acceptor_variant", "splice_donor_variant", "start_lost", "start_lost&splice_region_variant", "stop_gained", "stop_gained&splice_region_variant", "stop_lost", "stop_lost&splice_region_variant" and "stop_retained_variant”), and (2) filter on the MSC value, a gene-specific CADD threshold(28). These two filtering approaches resulted in a slightly higher TPR than our proposed strategy but lower TNR and lower precision (Fig 1B). Therefore, even in an exome analysis, the RAVA-FIRST filtering outperforms classical filtering strategies to select qualifying rare variants for RVAT.

Figure 1: TPR, TNR and precision of different filtering strategies on the whole Clinvar dataset or only Clinvar coding variants.

Finally, we investigated the performances of these different strategies on different classes of non-coding variants (S4 Fig). All the performances are lower than in the coding genome, especially the TPR that is much lower for strategies based on a fixed CADD threshold, highlighting the fact that CADD values are lower in the non-coding genome and adjusted CADD threshold may therefore be preferred. RAVA-FIRST filtering using region-dependant thresholds keeps the highest precision in the different classes of variants, except for UTR variants where a slight decrease of TNR and precision is observed.

Note however that these results may not be as accurate as those obtained on the coding regions as much fewer variants are included: 2,309, 4,048 and 617 for UTR, introns and intergenic variants respectively compared to 54,664 coding variants.

RAVA-FIRST burden test – Simulations

To validate the RAVA-FIRST burden test, we performed simulations under the null hypothesis and under different scenarios of association using data from the 1000 Genomes European populations (30) in the *LCT* gene. We simulated 1,000 controls and 1,000 cases using the simulations based on haplotypes implemented in the R package Ravages (19). A total of 201 variants was considered in the *LCT* gene. These variants were polymorphic in the European populations and rare variants were defined with a MAF lower than 1%. Two CADD regions overlap the *LCT* gene, R019233 and R019234, containing respectively 75 and 126 variants, both overlapping coding and regulatory categories.

Type I error

We first simulated data under the null hypothesis to verify that the RAVA-FIRST burden test maintains appropriate type I errors. We simulated two groups of 1,000 individuals in the R019234 CADD region without any genetic effect and we applied the classical WSS and the RAVA-FIRST WSS. Type I errors were computed using $5 \cdot 10^{6}$ simulations at three significance levels: $5 \cdot 10^{-2}$, 10^{-3} and $2.5 \cdot 10^{-6}$ (the usual threshold for whole exome rare variant association tests). The RAVA-FIRST WSS maintains good type I error levels at these different significance thresholds, similar to the ones obtained with the classical WSS (Table in S2 Table).

Power analysis

We then performed a power study based on simulations at two levels: at the level of the R019234 CADD region and at the level of the *LCT* gene. In both cases, we simulated 50% of causal variants randomly in the whole unit (scenarios S1 and S3), in the coding regions (scenarios S2A and S4A) or in

the regulatory regions (scenarios S2B and S4B). All the scenarios are summarised in Table 3. We compared the classical WSS to the RAVA-FIRST WSS using the gene or the two CADD regions as testing units. When CADD regions were used as testing units, analyses were performed for each of the two CADD regions and the minimum p-value was taken and multiplied by two to correct for multiple testing. A total of 1,000 replicates were simulated for each scenario and power was assessed at a genome-wide significance threshold of $2.5 \cdot 10^{-6}$.

Table 3: Scenarios of association simulated to assess the performance of the RAVA-FIRST burden test

	***LCT* gene**			
	R019233		**R019234**	
	Coding	Regulatory	Coding	Regulatory
S1			50%	
S2A			50%	0%
S2B			0%	50%
S3	50%			
S4A	50%	0%	50%	0%
S4B	0%	50%	0%	50%

Table 4 presents the power results obtained from this simulation study for both the classical WSS and the RAVA-FIRST WSS. Similar trends were observed between the two analyses, regardless if the simulations are performed at the scale of CADD regions or at the scale of the gene. When the causal variants were randomly sampled across the entire region (scenarios S1 and S3), the classical WSS with only one score for the entire region slightly outperformed the RAVA-FIRST method with sub-scores. Nevertheless, the loss of power for the latter was modest (less than 10%). By contrast, when causal variants were present only in the coding categories (scenarios S2A and S4A), which represent a small proportion of the entire region (approximately 15%), the RAVA-FIRST strategy was much more powerful than the classical WSS (approximately 50% gain in power). When causal variants were present in the regulatory categories only (scenarios S2B and S4B), both strategies showed similar

power. All these results highlight the gain of power using the RAVA-FIRST WSS when a cluster of causal variants is present within a functional category of the CADD region while maintaining good power levels when causal variants are spread all across the region. When comparing the simulations with causal variants sampled at the gene level or at the CADD region level, burden tests gathering variants within the corresponding testing units show, as expected, the highest levels of power. Nevertheless, the loss of power when using CADD regions as testing units instead of the entire gene is lower when causal variants are sampled across the entire gene (scenario S3) than the gain of power they present when causal variants are sampled within a specific CADD region (scenario S1). This is particularly true for the RAVA-FIRST WSS.

Table 4: Power at the genome-wide significance level of $2.5 \cdot 10^{-6}$ under the different simulation scenarios using either the classical WSS or the RAVA-FIRST WSS at the scale of either the entire gene or CADD regions

	By gene		**By CADD regions**	
	Classical WSS	RAVA-FIRST WSS	Classical WSS	RAVA-FIRST WSS
S1	0.409	0.370	0.782	0.701
S2A	0	0.431	0.002	0.602
S2B	0.408	0.404	0.689	0.706
S3	0.751	0.678	0.512	0.433
S4A	0.004	0.564	0.012	0.474
S4B	0.657	0.64	0.39	0.391

Applications

Ethics Statement

The MARTHA study was approved by its institutional ethics committee and informed written consent was obtained in accordance with the Declaration of Helsinki. Ethics approval were obtained from the "Departement santé de la direction générale de la recherche et de l'innovation du ministère" (Projects DC: 2008-880 and 09.576).

RAVA-FIRST analysis

RAVA-FIRST was used on whole genome sequence (WGS) data from patients affected by venous thromboembolism (VTE). VTE is a multifactorial disease with a strong genetic component (31). There exists a huge heterogeneity between patients in the age at first VTE event. To study the role of rare variants on VTE age of onset, WGS data were used from 200 individuals from the MARTHA cohort (32). These individuals were selected among patients with unprovoked VTE event who were previously genotyped for a genome-wide association study (33) and present no known genetic predisposing factor. Individuals were dichotomized based on the age at first VTE event either before 50 years of age (early-onset) or after (late-onset). The threshold of 50 years was chosen based on the results of recent studies (34) that hint toward a genetic heterogeneity between these two groups. A quality control (QC) of the sequencing data was performed using the program RAVAQ (https://gitlab.com/gmarenne/ravaq). After QC, 184 individuals were included for analysis with 127 presenting an early-onset VTE and 57 a late-onset VTE. Only variants passing all QC steps and with a MAF lower than 1% in the sample were considered in the association tests comparing early and late-onset groups. For these comparisons, rare variants were gathered either by CADD regions or by using the sliding windows procedure implemented in WGScan (18). Qualifying variants were selected based on CADD scores and using two filtering strategies: a fixed CADD threshold of 15 (as recommended by https://cadd.gs.washington.edu/info, version v1.4) or the RAVA-FIRST CADD region-specific filtering

(applied on adjusted scores). Association was tested using the WSS burden test. When the RAVA-FIRST filtering was used, the corresponding WSS test with sub-scores was applied. Table 5 shows the number of testing units and variants kept under each strategy. For all tests with CADD regions, only regions containing at least 5 rare variants were kept. WGScan was used with default parameters, i.e. with testing units of 5, 10, 15, 25 or 50 kb.

Table 5: Number of testing units and variants kept under the three strategies

	Testing units	Filtering	Number of testing units	Number of variants
WGScan Fixed CADD threshold	Sliding windows	MAF ≤ 1% CADD v1.4 ≥ 15	377,092	96,347
RAVA-FIRST units (CADD regions) Fixed CADD threshold	CADD regions	MAF ≤ 1% CADD v1.4 ≥ 15	10,389	96,294
RAVA-FIRST units (CADD regions) RAVA-FIRST filtering		MAF ≤ 1% Adjusted CADD ≥ median	95,690	3,641,502

QQ-plots for the WSS tests using those three strategies are shown in Fig 2. As expected, a lower significance threshold is required to reach genome-wide significance with the sliding window procedure due to the higher number of testing units. Accordingly, the computation time was much lower for the two analyses by CADD regions (6min when filtering based on a fixed CADD score threshold and 25min when using the region-specific CADD thresholds) than for the sliding windows procedure (47min). Our dataset contains less than 200 individuals, suggesting that the gain in computation time of CADD regions compared to sliding window procedures would be even greater in larger WGS datasets. No significant result was found when selecting variants with a CADD score greater than 15 using neither the sliding window strategies nor the CADD regions to gather rare variants, whereas one association reached borderline significance ($p = 6.41 \cdot 10^{-7}$) when using the RAVA-FIRST strategy.

Figure 2: QQ-plot of WSS analyses on VTE data using the three strategies of analysis. Early-onset patients (<50 years old) were compared to late-onset patients (≥50 years old).

This association maps to R126442, a CADD region of 21 kb on chromosome 18:66788277-66809402 that contains 31 rare variants after RAVA-FIRST filtering. In this region, none of the variants observed in VTE patients or in GnomAD achieved a CADD score above 15. This explains why the association could not have been detected by the two other strategies based on fixed CADD score ≥ 15. The median of CADD scores observed for GnomAD variants in this region is 1.44 and the adjusted CADD scores of selected variants range from 1.62 to 8.50. These observations emphasize the need to adapt thresholds depending on the genomic region under analysis. Interestingly, only early-onset VTE patients carry qualifying rare variants and have non-null WSS scores (Fig 3). Among early-onset patients, a trend is also observed for WSS scores to decrease with increasing age of onset.

Figure 3: WSS scores in the CADD region depending on the age at first VTE event. The dashed line corresponds to the age 50 discriminating early onset from late onset events.

The CADD region R126442 was then tested for association with 20 biological VTE biomarkers available in MARTHA patients: antithrombin, basophil, eosinophil, Factor VIII, Factor XI, fibrinogen, hematocrit, lymphocytes, mean corpuscular volume, mean platelet volume, monocytes, neutrophils, PAI-1, platelets count, protein C, protein S, prothrombin time, red blood cells count, von Willebrand Factor, and white blood cells count. For this, a linear regression model was used where adjustment was made on age at sampling and sex. At the Bonferonni threshold of 0.0025, one significant association ($p = 7.1 \cdot 10^{-4}$) was observed, VTE patients with a non-null WSS score exhibiting decreased haematocrit levels, a surrogate marker of red blood cells (Table in S3 Table). A similar trend ($p = 4.6 \cdot 10^{-3}$) was observed with red blood cell count.

We also investigated the association of the identified region with 376 plasma protein antibodies that were selected to be involved in thrombosis-related processes and that have been previously profiled in MARTHA (32,35). Regression analysis were conducted on log transformed values of antibodies and

were adjusted for age, sex, and three internal control antibodies. In order to handle the correlation between measured protein antibodies, we used the Li and Ji method (36) to estimate the number of effective independent tests. This number, calculated to be 163, was then used to define a Bonferroni threshold for declaring study-wide statistical significance. While not reaching the study-wise significance level of $p = 3.1 \cdot 10^{-4}$ after correction for multiple testing, it is worth noting that the two proteins that exhibited the strongest significance with marginal association at $p < 0.001$, procalcitonin tagged by the HPA043700 antibody ($p = 7.2 \cdot 10^{-4}$) and PDPK1 tagged by HPA035199 ($p = 7.5 \cdot 10^{-4}$), have been both proposed to be involved in red blood cell biology (37,38).

According to ENCODE data, the R1246442 CADD region overlaps “intergenic” and “regulatory” categories with one distant enhancer-like signature. To describe this region further, we looked at TADs positions in https://dna.cs.miami.edu/TADKB/brows.php in HUVEC and HMEC cell lines, two cell types known to be relevant for VTE pathophysiology. We found that the CADD region is included into the topological associated domains (TADs) 18:66450000-68150000. By studying TADs described by Lieberman-Aiden et al. 2009 in other cell lines such as KBM7, K562 or GM12878, we retrieved a TAD with similar positions, giving additional evidence for the presence of this TAD around the CADD region associated with early-onset patients. We then explored this TAD region for the presence of candidate VTE genes whose regulation could be influenced by the enhancer region that maps our R1246442 region. Using the UCSC genome browser (40) integrating information about interactions between GeneHancer regulatory elements and genes expression (see S5 Fig), we identified *CD226* as a strong biological candidate. *CD226* codes for a glycoprotein expressed at the surface of several types of cells, including blood cell, and several studies have shown that it was associated with vascular endothelial dysfunction (41–43). Genetic variants in *CD226* have also been found associated with several blood cell traits including platelets, white blood cells (e.g. neutrophil, eosinophil) (44) and reticulocyte counts (45), another red blood cell biomarker.

Discussion

Even though whole genome sequencing data are now more often available on cases and controls, rare variant association tests (RVAT) usually remain restricted to the coding part of the genome. This is explained by the lack of tools to explore rare variant associations outside genes (11). Indeed, RVAT requires the definition of testing units that are easily defined through genes in coding regions and the selection in these regions of the most functionally-relevant variants. This is also easier in the coding genome as most prediction tools were developed and tested through the effects of variants on encoded proteins. In the non-coding genome, testing units can be defined based on functional elements such as enhancers or silencers, or through the use of sliding window procedures. The first solution prevents RVAT from being applied to all rare variants in the genome as biological units are not defined over the entire genome. The second strategy with sliding windows results in a large number of tests and the need to adjust p-values to take into account the multiple correlated tests performed. In this work, we propose an entire new strategy of analysis of rare variants in the coding and the non-coding genome, RAVA-FIRST, which is composed of three steps. Firstly, RAVA-FIRST proposes some new testing units to gather rare variants, the so-called “CADD regions” that we defined over the entire genome based on CADD scores of variants observed in GnomAD. These CADD regions are large enough to include a sufficient number of rare variants to allow RVAT. They tend to preserve functional elements that, for a majority of them, are not split into several CADD regions. Secondly, RAVA-FIRST filters variants based on region-specific adjusted CADD thresholds that allow to select the best candidate variants within each region. This filtering approach was found to be more efficient than traditional approaches to discriminate between benign and pathogenic variants within a set of variants. Indeed, our benchmarking study using a set of Clinvar variants showed that the other filtering strategies we considered were good at identifying true causal variants (true positive rates were high) but bad at finding the non-causal variants (true negative rates were low). Both true positive and true negative rates are important to achieve a high percentage of causal variants within testing units, this

415 percentage being the main driver of power in RVAT, especially in burden tests (2,3,7). Thus, the RAVA-
416 FIRST filtering strategy is expected to result in an appreciable increase of power as compared to
417 classically used strategies. Indeed, RAVA-FIRST enables to keep the most important functional variants
418 within coding, regulatory and intergenic categories of the genome by adapting CADD score threshold
419 to the genomic context. Finally, RAVA-FIRST includes a burden test that integrates information on
420 genomic categories in the regression and that, coupled with the region-specific filtering, leads to a
421 better detection of causal variants, should they cluster in one of these genomic categories only. We
422 also showed through simulations that good power levels were maintained using RAVA-FIRST burden
423 test when causal variants were randomly sampled.

424 RAVA-FIRST was applied on real WGS data from VTE patients where an accumulation of rare variants
425 in patients with early-onset events was investigated. We did not detect any significant signal using the
426 sliding window procedure or CADD regions when qualifying rare variants were selected based on a
427 fixed CADD threshold. However, we detected an association signal using both the grouping and filtering
428 of rare variants proposed in RAVA-FIRST. The associated CADD region is intergenic, contains a
429 predicted enhancer and is surrounded by a TAD containing 5 genes including *CD226*, a strong candidate
430 for blood cell traits that are new well recognized to be key players in VTE physiopathology (31). All rare
431 variants in this region present low CADD scores and were not even included in analyses based on a fix
432 CADD threshold, highlighting the importance of taking into account the genetic context to detect the
433 most important predicted functional variants within each CADD region. These 31 rare variants are
434 exclusively observed in early-onset cases. Fourteen of these variants are absent from GnomAD, and 10
435 of the 17 remaining variants have a lower frequency in GnomAD population than in our sample. This
436 reinforces the value of the association signal in this CADD region, although it should be further
437 described and validated using functional experiments. Preliminary investigations that need to be
438 further explored, at both experimental and epidemiological levels, strongly suggest that this region is
439 associated with several inflammatory markers impaired in anaemia of inflammation (38,46) and in
440 platelets, both mechanisms being involved in thrombotic processes (47).

Some limits can be pointed out on our RAVA-FIRST approach. Firstly, the definition of CADD regions relies on the GnomAD population and on the adjusted CADD threshold. We chose to use the whole GnomAD dataset but it could be of interest to select some of the populations to be more specific. It has for example been suggested that different expression patterns could be found between different populations (48). Nevertheless, in classical exome analyses, rare variants are mostly filtered based on the maximum frequency observed among multiple populations. Furthermore, CADD regions are not defined for low-covered and non-sequenced genomic regions in GnomAD and their definition could therefore be improved in the future. Concerning the definition of the genomic categories, we decided to include all genomic elements directly implicated into regulatory functions to define the regulatory regions of the genome, but we did not include silencers or lncRNA for example. However, the choice of elements to include as the regulatory category will only impact the adjusted CADD scores that are similar between regulatory and intergenic regions, and won't therefore have a huge impact on CADD regions definition. As an example, using DECRES (49) to predict enhancers and promoters instead of SCREEN results in a very high correlation between the definition of CADD regions, 80% of them being identical.

On the other hand, the pre-definition of regions in the whole genome offers several advantages, including the region-specific filtering mentioned before. In addition, the newly defined CADD regions can be used in existing software that require regions as input parameters (50,51), enabling to apply a wide variety of RVAT available in those programs to the whole genome. Especially, Bayesian methods which have been shown to be of great promise in the analysis and filtering of rare variants (52,53) could be applied beyond genes by using CADD regions.

To our knowledge, CADD regions represent predefined testing units for RVAT that cover the highest proportion of the genome. These regions have been made publicly available (cf "Data availability" section below). CADD regions are part of a whole new strategy of rare variant analysis in the whole genome, RAVA-FIRST, that further benefits from the integration of functional information both for the

466 filtering of rare variants and their analysis with burden tests. RAVA-FIRST has been implemented in the
467 package R Ravages available in the CRAN and on Github, offering an easy and straightforward tool to
468 perform RVAT in the whole genome. We believe that our developments will help researchers to
469 explore the role of genome-wide rare variants in complex diseases. Firstly, through the redefinition of
470 testing units in the coding genome where cluster of causal variants can be found within genes and
471 retrieved using CADD regions (10). Secondly, through the study of non-coding variants, especially
472 intergenic ones, which are currently often excluded from the analysis. Going beyond the gene and the
473 consequences on proteins, RAVA-FIRST will help for a better understanding of biological mechanisms
474 behind complex diseases.

475 Data availability

476 The files containing the positions of CADD regions, the positions of genomic categories and the
477 adjusted CADD scores are available at https://lysine.univ-brest.fr/RAVA-FIRST/. All the functions
478 needed for RAVA-FIRST to annotate, group, filter and analyse rare variants have been implemented in
479 the package R Ravages (https://cran.r-project.org/web/packages/Ravages/,
480 https://github.com/genostats/Ravages) which directly downloads the files from https://lysine.univ-
481 brest.fr/RAVA-FIRST/.

482 Information about the CADD region R126442 that was found associated with VTE age at first event is
483 available in the Supporting Information File 2. Information about individuals (WSS score, age and sex)
484 and variants (position, adjusted CADD score and weight in WSS) are given.

References

1. Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. PLoS genetics. 2009;5(2):e1000384.

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

3. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. Am J Hum Genet. 2014 Jul 3;95(1):5–23.

4. Bellenguez C, Charbonnier C, Grenier-Boley B, Quenez O, Le Guennec K, Nicolas G, 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 Nov;59:220.e1-220.e9.

5. Shaffer JR, LeClair J, Carlson JC, Feingold E, Buxó CJ, Christensen K, et al. Association of low-frequency genetic variants in regulatory regions with nonsyndromic orofacial clefts. American Journal of Medical Genetics Part A. 2019 Mar;179(3):467–74.

6. Wang Q, Dhindsa RS, Carss K, Harper AR, Nag A, Tachmazidou I, et al. Rare variant contribution to human disease in 281,104 UK Biobank exomes. Nature [Internet]. 2021 Aug 10 [cited 2021 Aug 12]; Available from: https://www.nature.com/articles/s41586-021-03855-y

7. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet. 2008 Sep;83(3):311–21.

8. Bis JC, Jian X, Kunkle BW, Chen Y, Hamilton-Nelson KL, Bush WS, et al. Whole exome sequencing study identifies novel rare and common Alzheimer's-Associated variants involved in immune response and transcriptional regulation. Mol Psychiatry. 2018 Aug 14;

9. Cirulli ET, White S, Read RW, Elhanan G, Metcalf WJ, Tanudjaja F, et al. Genome-wide rare variant analysis for thousands of phenotypes in over 70,000 exomes from two cohorts. Nat Commun [Internet]. 2020 Jan 28 [cited 2020 May 18];11. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6987107/

10. Guey S, Kraemer M, Hervé D, Ludwig T, Kossorotoff M, Bergametti F, et al. Rare RNF213 variants in the C-terminal region encompassing the RING-finger domain are associated with moyamoya angiopathy in Caucasians. Eur J Hum Genet. 2017;25(8):995–1003.

11. Bocher O, Génin E. Rare variant association testing in the non-coding genome. Hum Genet [Internet]. 2020 Jun 4 [cited 2020 Jun 8]; Available from: http://link.springer.com/10.1007/s00439-020-02190-y

12. UK10K Consortium, Walter K, Min JL, Huang J, Crooks L, Memari Y, et al. The UK10K project identifies rare variants in health and disease. Nature. 2015 Oct 1;526(7571):82–90.

13. Zhang F, Lupski JR. Non-coding genetic variants in human disease. Hum Mol Genet. 2015 Oct 15;24(R1):R102-110.

14. Castel SE, Cervera A, Mohammadi P, Aguet F, Reverter F, Wolman A, et al. Modified penetrance of coding variants by cis-regulatory variation contributes to disease risk. Nature Genetics. 2018 Sep;50(9):1327–34.

15. Morrison AC, Huang Z, Yu B, Metcalf G, Liu X, Ballantyne C, et al. Practical Approaches for Whole-Genome Sequence Analysis of Heart- and Blood-Related Traits. The American Journal of Human Genetics. 2017 Feb;100(2):205–15.

16. Cochran JN, Geier EG, Bonham LW, Newberry JS, Amaral MD, Thompson ML, et al. Non-coding and Loss-of-Function Coding Variants in TET2 are Associated with Multiple Neurodegenerative Diseases. Am J Hum Genet. 2020 May 7;106(5):632–45.

17. Li Z, Li X, Liu Y, Shen J, Chen H, Zhou H, et al. Dynamic Scan Procedure for Detecting Rare-Variant Association Regions in Whole-Genome Sequencing Studies. The American Journal of Human Genetics. 2019 May;104(5):802–14.

18. He Z, Xu B, Buxbaum J, Ionita-Laza I. A genome-wide scan statistic framework for whole-genome sequence data analysis. Nature Communications. 2019 Jul 9;10(1):1–11.

19. Bocher O, Marenne G, Pierre AS, Ludwig TE, Guey S, Tournier-Lasserve E, et al. Rare variant association testing for multicategory phenotype. Genetic Epidemiology. 2019;43(6):646–56.

20. Bocher O, Marenne G, Tournier-Lasserve E, FREX Consortium, Génin E, Perdry H. Extension of SKAT to multi-category phenotypes through a geometrical interpretation. Eur J Hum Genet [Internet]. 2021 Jan 14 [cited 2021 Jan 15]; Available from: http://www.nature.com/articles/s41431-020-00792-8

21. Havrilla JM, Pedersen BS, Layer RM, Quinlan AR. A map of constrained coding regions in the human genome. Nature Genetics. 2019 Jan;51(1):88–95.

22. Genome Aggregation Database Consortium, Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020 May;581(7809):434–43.

23. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Research. 2019 Jan 8;47(D1):D886–94.

24. Nishizaki SS, Boyle AP. Mining the Unknown: Assigning Function to Noncoding Single Nucleotide Polymorphisms. Trends in Genetics. 2017 Jan;33(1):34–45.

25. Pujar S, O'Leary NA, Farrell CM, Loveland JE, Mudge JM, Wallin C, et al. Consensus coding sequence (CCDS) database: a standardized set of human and mouse protein-coding regions supported by expert curation. Nucleic Acids Res. 2018 04;46(D1):D221–8.

26. Barrett LW, Fletcher S, Wilton SD. Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements. Cell Mol Life Sci. 2012 Nov;69(21):3613–34.

27. Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N, Adrian J, et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature. 2020 Jul;583(7818):699–710.

28. Itan Y, Shang L, Boisson B, Ciancanelli MJ, Markle JG, Martinez-Barricarte R, et al. The mutation significance cutoff: gene-level thresholds for variant predictions. Nat Methods. 2016 Feb;13(2):109–10.

29. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018 Jan 4;46(D1):D1062–7.

30. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015 Oct 1;526(7571):68–74.

31. Lindström S, Wang L, Smith EN, Gordon W, van Hylckama Vlieg A, de Andrade M, et al. Genomic and transcriptomic association studies identify 16 novel susceptibility loci for venous thromboembolism. Blood. 2019 Nov 7;134(19):1645–57.

32. Razzaq M, Iglesias MJ, Ibrahim-Kosta M, Goumidi L, Soukarieh O, Proust C, et al. An artificial neural network approach integrating plasma proteomics and genetic data identifies PLXNA4 as a new susceptibility locus for pulmonary embolism. Sci Rep. 2021 Jul 7;11(1):14015.

33. Germain M, Saut N, Greliche N, Dina C, Lambert J-C, Perret C, et al. Genetics of Venous Thrombosis: Insights from a New Genome Wide Association Study. PLOS ONE. 2011 Sep 27;6(9):e25581.

34. Roupie A-L, Dossier A, Goulenok T, Perozziello A, Papo T, Sacre K. First venous thromboembolism in admitted patients younger than 50years old. European Journal of Internal Medicine. 2016 Oct 1;34:e18–20.

35. Razzaq M, Goumidi L, Iglesias M-J, Munsch G, Bruzelius M, Ibrahim-Kosta M, et al. Explainable Artificial Neural Network for Recurrent Venous Thromboembolism Based on Plasma Proteomics. In: Cinquemani E, Paulevé L, editors. Computational Methods in Systems Biology. Cham: Springer International Publishing; 2021. p. 108–21. (Lecture Notes in Computer Science).

36. Li J, Ji L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. Heredity. 2005 Sep;95(3):221–7.

37. Cokic VP, Bhattacharya B, Beleslin-Cokic BB, Noguchi CT, Puri RK, Schechter AN. JAK-STAT and AKT pathway-coupled genes in erythroid progenitor cells through ontogeny. J Transl Med. 2012 Jun 7;10:116.

38. Weiss G, Ganz T, Goodnough LT. Anemia of inflammation. Blood. 2019 Jan 3;133(1):40–50.

39. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long range interactions reveals folding principles of the human genome. Science. 2009 Oct 9;326(5950):289–93.

40. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The Human Genome Browser at UCSC. Genome Res. 2002 Jan 6;12(6):996–1006.

41. Chen L, Xie X, Zhang X, Jia W, Jian J, Song C, et al. The expression, regulation and adhesion function of a novel CD molecule, CD226, on human endothelial cells. Life Sci. 2003 Sep 19;73(18):2373–82.

42. Kojima H, Kanada H, Shimizu S, Kasama E, Shibuya K, Nakauchi H, et al. CD226 mediates platelet and megakaryocytic cell adhesion to vascular endothelial cells. J Biol Chem. 2003 Sep 19;278(38):36748–53.

598 43. Zhou S, Xie J, Yu C, Feng Z, Cheng K, Ma J, et al. CD226 deficiency promotes glutaminolysis and
599 alleviates mitochondria damage in vascular endothelial cells under hemorrhagic shock. FASEB J.
600 2021 Nov;35(11):e21998.

601 44. Chen M-H, Raffield LM, Mousas A, Sakaue S, Huffman JE, Moscati A, et al. Trans-ethnic and
602 Ancestry-Specific Blood-Cell Genetics in 746,667 Individuals from 5 Global Populations. Cell.
603 2020 Sep 3;182(5):1198-1213.e14.

604 45. Vuckovic D, Bao EL, Akbari P, Lareau CA, Mousas A, Jiang T, et al. The Polygenic and Monogenic
605 Basis of Blood Traits and Diseases. Cell. 2020 Sep 3;182(5):1214-1231.e11.

606 46. Nemeth E, Ganz T. Anemia of inflammation. Hematol Oncol Clin North Am. 2014
607 Aug;28(4):671–81, vi.

608 47. Wagner DD, Burger PC. Platelets in inflammation and thrombosis. Arterioscler Thromb Vasc
609 Biol. 2003 Dec;23(12):2131–7.

610 48. Halachev M, Meynert A, Taylor MS, Vitart V, Kerr SM, Klaric L, et al. Increased ultra-rare variant
611 load in an isolated Scottish population impacts exonic and regulatory regions. PLoS Genet. 2019
612 Nov;15(11):e1008480.

613 49. Li Y, Shi W, Wasserman WW. Genome-wide prediction of cis-regulatory regions using
614 supervised deep learning methods. BMC Bioinformatics. 2018 May 31;19(1):202.

615 50. Zhan X, Hu Y, Li B, Abecasis GR, Liu DJ. RVTESTS: an efficient and comprehensive tool for rare
616 variant association analysis using sequence data. Bioinformatics. 2016 May 1;32(9):1423–6.

617 51. Baskurt Z, Mastromatteo S, Gong J, Wintle RF, Scherer SW, Strug LJ. VikNGS: a C++ variant
618 integration kit for next generation sequencing association analysis. Bioinformatics. 2020 Feb
619 15;36(4):1283–5.

620 52. Quintana MA, Berstein JL, Thomas DC, Conti DV. Incorporating model uncertainty in detecting
621 rare variants: the Bayesian risk index. Genetic Epidemiology. 2011 Nov;35(7):638–49.

622 53. Greene D, Richardson S, Turro E. A Fast Association Test for Identifying Pathogenic Variants
623 Involved in Rare Diseases. The American Journal of Human Genetics. 2017 Jul;101(1):104–14.

624

625

Supporting Information captions

S1 Fig. Steps performed in RAVA-FIRST.

S2 Fig. Definition of CADD regions and removal of low-covered and non-sequenced regions in GnomAD.

S3 Fig. Percentage of CADD regions (≥1kbp) overlapping each of the three genomic categories.

S4 Fig. TPR, TNR and precision of different filtering strategies on Clinvar non-coding variants (UTR, introns or intergenic regions).

S5 Fig. Screenshot of the TAD 18:66450000-68150000 in the UCSC genome browser containing the CADD region R126442 and a potential enhancer regulating the CD226 gene, a candidate gene in VTE.

S1 Table. Sources used to get genomic elements for comparisons with CADD regions.

S2 Table. Type I error of the classical WSS and the RAVA-FIRST WSS using $5 \cdot 10^6$ simulations under the null hypothesis.

S3 Table. Characteristics of the studied VTE sample. Mean (Standard Deviation) for quantitative variables. Count (%) for qualitative variables.

S1 File. Details about the RAVA-RIST method and its evaluation.

S2 File. Information on the CADD region R126442 associated with VTE age at onset. Information about individuals (WSS score, age and sex) and variants (position, adjusted CADD score and weight in WSS) are given.

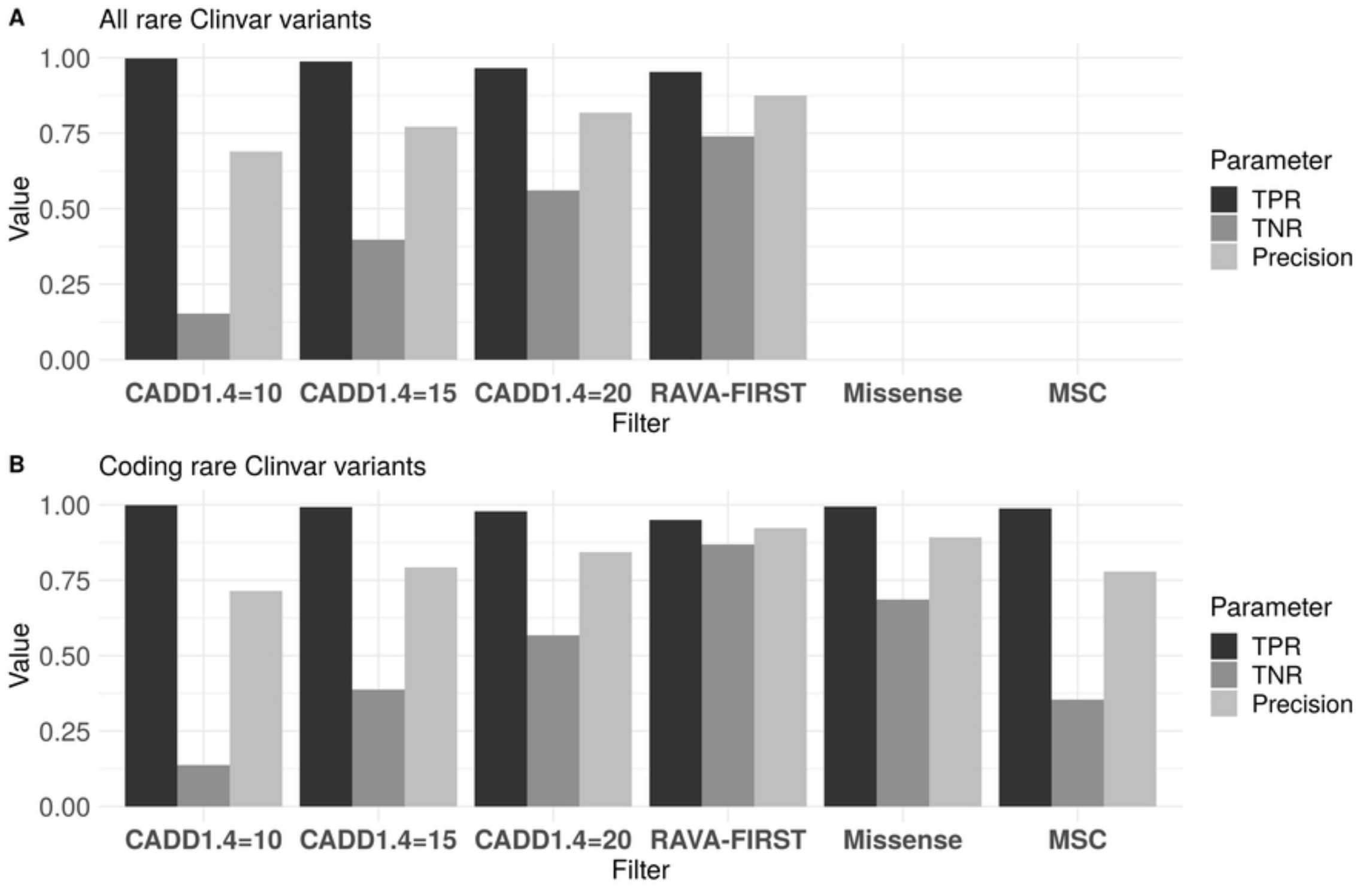
A
All rare Clinvar variants
1.00
0.75
0.50
0.25
0.00
Value
CADD1.4=10
CADD1.4=15
CADD1.4=20
RAVA-FIRST
Missense
MSC
Filter
Parameter
TPR
TNR
Precision
B
Coding rare Clinvar variants
1.00
0.75
0.50
0.25
0.00
Value
CADD1.4=10
CADD1.4=15
CADD1.4=20
RAVA-FIRST
Missense
MSC
Filter
Parameter
TPR
TNR
Precision

Figure 1

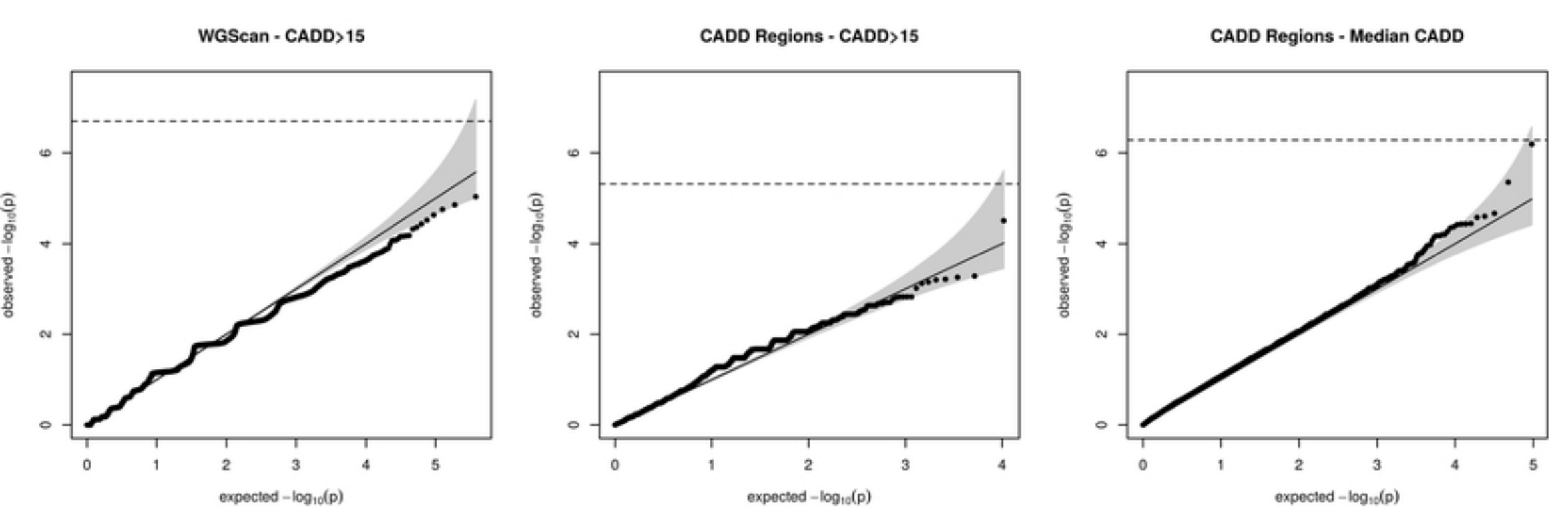
WGScan - CADD>15
CADD Regions - CADD>15
CADD Regions - Median CADD
observed −log10(p)
expected −log10(p)

Figure 2

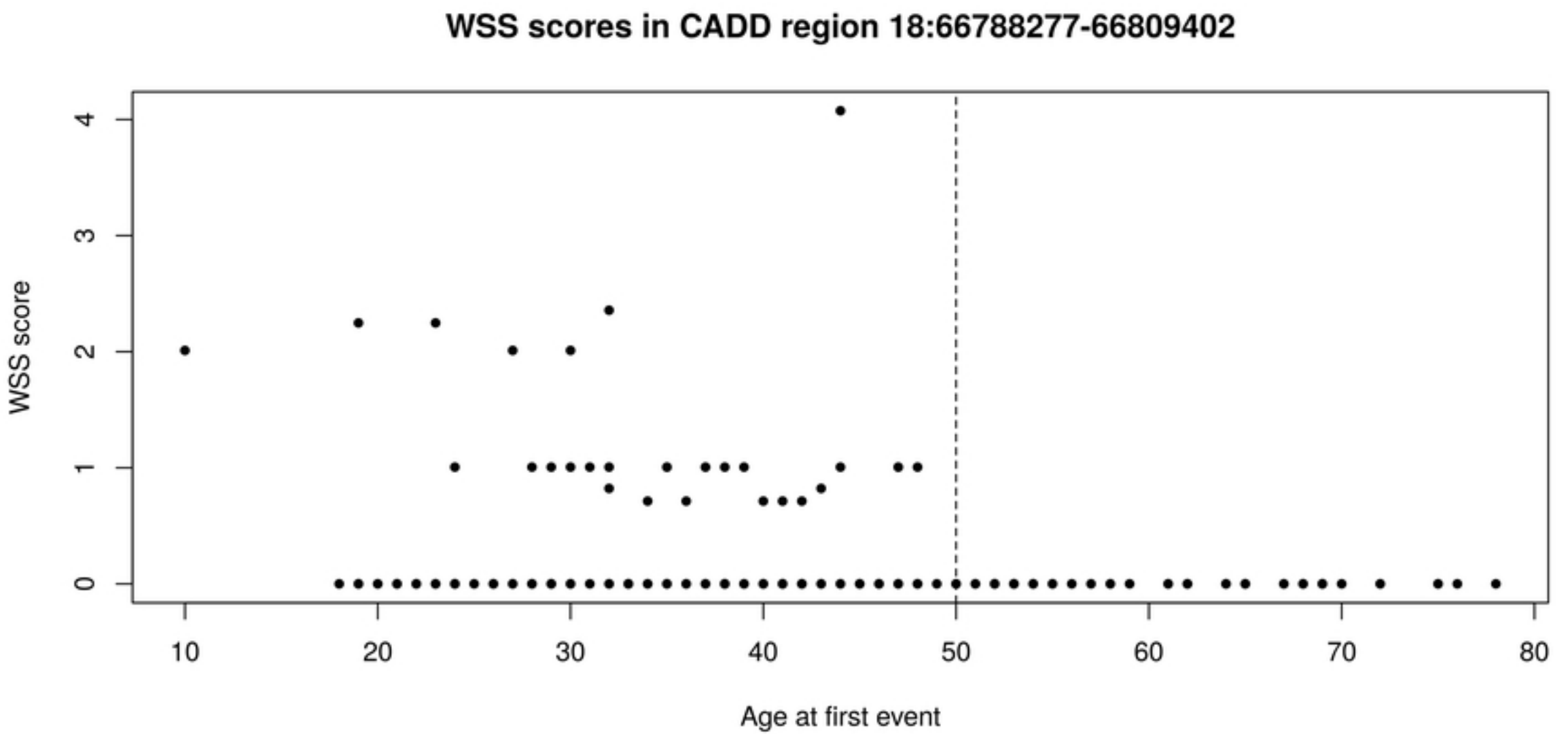
WSS scores in CADD region 18:66788277-66809402
WSS score
0
1
2
3
4
10
20
30
40
50
60
70
80
Age at first event

Figure 3