

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

4 Ozvan Bocher<sup>1,\*</sup>, Thomas E. Ludwig<sup>1,2</sup>, Gaëlle Marenne<sup>1</sup>, Jean-François Deleuze<sup>3</sup>, Suryakant

5 Suryakant<sup>4</sup>, Jacob Odeberg<sup>5,6</sup>, Pierre-Emmanuel Morange<sup>7</sup>, David-Alexandre Trégouët<sup>4</sup>, Hervé

6 Perdry<sup>8</sup>, Emmanuelle Génin<sup>1,2</sup>

7 <sup>1</sup>Univ Brest, Inserm, EFS, UMR 1078, GGB, F-29200 Brest, France

8 <sup>2</sup>CHU Brest, Brest, France

9 <sup>3</sup>Centre National de Recherche en Génomique Humaine CNRGH, Institut de Biologie François Jacob,

10 Université Paris Saclay, CEA, 2 rue Gaston Crémieux CP 5721, 91057 Evry, France

11 <sup>4</sup>University of Bordeaux, Inserm, Bordeaux Population Health Research Center, team ELEANOR, UMR

12 1219, 33000 Bordeaux, France

13 <sup>5</sup>Science for Life Laboratory, Department of Protein Science, CBH, KTH Royal Institute of Technology,

14 Stockholm, Sweden

15 <sup>6</sup>Department of Clinical Medicine, Faculty of Health Science, The Arctic University of Tromsö, Tromsö,

16 Norway

17 <sup>7</sup>Aix Marseille Univ, INSERM, INRAE, C2VN, Marseille, France

18 <sup>8</sup>CESP Inserm, U1018, UFR Médecine, Univ Paris-Sud, Université Paris-Saclay, Villejuif, France

19 \*Corresponding author

20 Email: [bocherozvan@gmail.com](mailto:bocherozvan@gmail.com) (OB)

## 21 Abstract

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

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

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

42

43

## 44 Author Summary

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

56

## 57 Introduction

58 With advance in sequencing technologies, it is now possible to explore the role of rare genetic variants  
59 in complex diseases. Different rare variant association tests (RVAT) have been developed that gather  
60 rare variants into testing units and compare rare variant content in these testing units between cases  
61 and controls (1–3). While the impact of rare variants has already been shown in several complex  
62 diseases (4–6), RVAT face two key challenges: (i) the definition of the testing units and (ii) the selection  
63 of the qualifying rare variants to include in these units. The proportion of causal variants in the testing  
64 units being a major driver of power, especially for burden tests, it is indeed important to ensure that  
65 qualifying variants are enriched in variants likely to have some functional impact (3,7). When exome  
66 analyses are undertaken, rare variants are most often grouped by genes and included in the analysis  
67 depending on their impact on the corresponding protein (8,9). Nevertheless, the gene definition is not  
68 always optimal as differences in rare variants burden between cases and controls could sometimes  
69 only be found in a sub-region of a gene. This is for example the case in the *RNF213* gene where an  
70 enrichment in rare variants located in the C-terminal region is found in Moyamoya cases (10). Defining  
71 testing units and qualifying variants is also much more challenging in the non-coding genome due to  
72 the lack of defined genomic elements and the higher difficulty to predict the functional impact of non-  
73 coding variants (11). It is yet a question of interest as several studies have shown the importance of  
74 rare non-coding variants in the development of complex diseases (12–14). Functional elements such  
75 as enhancers or promoters can be used as testing units (5,15,16) but they prevent the analysis of all  
76 rare variants in the genome and can be too small to get a sufficient number of rare variants for  
77 association analysis. On the other hand, sliding windows procedures such as SCAN-G (17) or WGSCAN  
78 (18) can be used to test for association over the whole genome. Nevertheless, they present several  
79 limits including the window definition that is arbitrary and blind to biological information, the high  
80 number of tests and the associated computation time. With overlapping windows, there is also a  
81 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

## 104 Description of the Method

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

## 109 Testing units in RAVA-FIRST: the CADD regions

110 Following Havrilla et al. (2019) (21), we seek to identify some genomic regions that were significantly  
111 depleted in functional variants to use them as testing units in RVAT. For that purpose, Havrilla et al.  
112 (2019) defined “constrained coding regions” (CCR) as exonic regions where no important functional  
113 variation (defined as being at least missense) was found in the general population of GnomAD (22). In  
114 our experience, two limits prevent the direct use of CCR as testing units in the whole genome: they are  
115 too small to gather a sufficient number of rare variants (224 bp being the maximum length of a CCR)  
116 and their definition relies on the consequence of the variants on the translated protein, not available  
117 in the non-coding genome. We therefore decided to expand the proposed approach by estimating the  
118 functionality of variants through CADD scores (23). CADD scores were chosen because of their  
119 availability for every substitution in the genome and because they rank well in the comparison test of  
120 functional annotation tools (24).

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

128 the definition of a large number of regulatory elements in diverse cell types (27). Finally, intergenic  
129 regions correspond to all regions not being described as coding or regulatory regions, representing  
130 54.5% of the genome. More details are given in the Supporting Information.

131 Adjusted CADD scores were used to select the variants that will bound the “CADD regions”. First, we  
132 selected the variants with an adjusted CADD score greater than 20, that is the top 1% of variants with  
133 the highest predicted functional impact within each of the three genomic categories. Then, among  
134 those variants, only the ones observed at least two times in GnomAD r2.0.1 genomes were further  
135 selected and used as boundaries of CADD regions. For CADD regions to be used as testing units in RVAT,  
136 they need to be large enough to contain several rare variants. Contiguous small regions of less than  
137 10 kb were therefore grouped together to form clusters of variants with high adjusted CADD scores.  
138 Non-sequenced regions and low-covered regions in GnomAD containing potential important  
139 functional variants were excluded from CADD regions, leading to gaps within CADD regions of at least  
140 one base pair (i.e. no CADD region overlap them to avoid artificially long regions due to a lack of  
141 variants in GnomAD). Finally, CADD regions are only defined for regions where CADD scores are  
142 available (removing among others centromeres and telomeres). Note that CADD regions can overlap  
143 different genomic categories (coding, regulatory or intergenic). More details about the steps and  
144 parameters used for the definition of CADD regions are presented in the Supporting Information.

145

## 146 The RAVA-FIRST filtering strategy

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

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

158

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

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

$$163 \ln \frac{P(Y_j = 1)}{P(Y_j = 0)} = \beta_0 + \beta_{Cov} X_{Cov} + \beta_G X_G$$

164 With  $Y_j$  the vector of phenotypes for the  $n$  individuals: 0 for the group of controls and 1 for the group  
165 of cases.  $\beta_0$  represents the intercept of the model and  $X_{Cov}$  the matrix of covariates (if any) with their  
166 associated effect,  $\beta_{Cov}$ .  $\beta_G$  corresponds to the estimated effect of the burden  $X_G$ , computed for  
167 example using WSS (1) which corresponds to a weighted sum of rare alleles based on their frequency,  
168 the rarest alleles having the highest weights.

169 To take into account functional information, we integrated a sub-score for each genomic category into  
170 the regression model, similarly to the analysis of rare and frequent variants proposed by Li and Leal  
171 (2008) (7):

$$172 \ln \frac{P(Y_j = 1)}{P(Y_j = 0)} = \beta_0 + \beta_{Cov} X_{Cov} + \sum_{G=\{cod;reg;inter\}} \beta_G X_G$$

173

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

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

182

## 183 Verification and Comparison

### 184 Statistics on CADD regions and comparison with genomic elements

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

194

195 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

196

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

207 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%

208

## 209 Performance of RAVA-FIRST filtering based on adjusted CADD scores

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

218 increases with the increasing CADD score threshold which is expected as less variants, and therefore  
219 less benign variants, are included in the analysis. The RAVA-FIRST filtering shows the highest TNR and  
220 the highest precision. While the TPR value is extremely important to select the most probable causal  
221 variants in RVAT, it is also important to have a high TNR value, otherwise the signal will be diluted by  
222 a high proportion of non-causal variants. The precision value summarises the TPR and TNR parameters  
223 and therefore, to a certain extent, is representative of the percentage of causal variants among  
224 selected variants. Therefore, we show that the RAVA-FIRST filtering strategy is the most accurate to  
225 select qualifying rare variants for RVAT. Focusing on the coding genome, we also compared the  
226 performance of RAVA-FIRST filtering approach against two others approaches classically used on genes  
227 as testing units: (1) filter for variants with a functional impact expected to change the protein  
228 ("missense\_variant", "missense\_variant&splice\_region\_variant", "splice\_acceptor\_variant",  
229 "splice\_donor\_variant", "start\_lost", "start\_lost&splice\_region\_variant", "stop\_gained",  
230 "stop\_gained&splice\_region\_variant", "stop\_lost", "stop\_lost&splice\_region\_variant" and  
231 "stop\_retained\_variant"), and (2) filter on the MSC value, a gene-specific CADD threshold(28). These  
232 two filtering approaches resulted in a slightly higher TPR than our proposed strategy but lower TNR  
233 and lower precision (Fig 1B). Therefore, even in an exome analysis, the RAVA-FIRST filtering  
234 outperforms classical filtering strategies to select qualifying rare variants for RVAT.

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

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

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

246

## 247 RAVA-FIRST burden test – Simulations

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

### 255 Type I error

256 We first simulated data under the null hypothesis to verify that the RAVA-FIRST burden test maintains  
257 appropriate type I errors. We simulated two groups of 1,000 individuals in the R019234 CADD region  
258 without any genetic effect and we applied the classical WSS and the RAVA-FIRST WSS. Type I errors  
259 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  
260 threshold for whole exome rare variant association tests). The RAVA-FIRST WSS maintains good type I  
261 error levels at these different significance thresholds, similar to the ones obtained with the classical  
262 WSS (Table in S2 Table).

### 263 Power analysis

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

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

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

	<b><i>LCT</i> gene</b>			
	<b>R019233</b>		<b>R019234</b>	
	Coding	Regulatory	Coding	Regulatory
S1			50%	
S2A			50%	0%
S2B			0%	50%
S3	50%			
S4A	50%	0%	50%	0%
S4B	0%	50%	0%	50%

274  
275 Table 4 presents the power results obtained from this simulation study for both the classical WSS and  
276 the RAVA-FIRST WSS. Similar trends were observed between the two analyses, regardless if the  
277 simulations are performed at the scale of CADD regions or at the scale of the gene. When the causal  
278 variants were randomly sampled across the entire region (scenarios S1 and S3), the classical WSS with  
279 only one score for the entire region slightly outperformed the RAVA-FIRST method with sub-scores.  
280 Nevertheless, the loss of power for the latter was modest (less than 10%). By contrast, when causal  
281 variants were present only in the coding categories (scenarios S2A and S4A), which represent a small  
282 proportion of the entire region (approximately 15%), the RAVA-FIRST strategy was much more  
283 powerful than the classical WSS (approximately 50% gain in power). When causal variants were  
284 present in the regulatory categories only (scenarios S2B and S4B), both strategies showed similar

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

294 Table 4: Power at the genome-wide significance level of  $2.5 \cdot 10^{-6}$  under the different simulation  
295 scenarios using either the classical WSS or the RAVA-FIRST WSS at the scale of either the entire gene  
296 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

297

## 298 Applications

### 299 Ethics Statement

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

### 304 RAVA-FIRST analysis

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

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

327 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 $\leq$ 1% CADD v1.4 $\geq$ 15	377,092	96,347
RAVA-FIRST units (CADD regions) Fixed CADD threshold	CADD regions	MAF $\leq$ 1% CADD v1.4 $\geq$ 15	10,389	96,294
RAVA-FIRST units (CADD regions) RAVA-FIRST filtering		MAF $\leq$ 1% Adjusted CADD $\geq$ median	95,690	3,641,502

328

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

340 **Figure 2: QQ-plot of WSS analyses on VTE data using the three strategies of analysis.** Early-onset

341 patients (<50 years old) were compared to late-onset patients (≥50 years old).

342 This association maps to R126442, a CADD region of 21 kb on chromosome 18:66788277-66809402

343 that contains 31 rare variants after RAVA-FIRST filtering. In this region, none of the variants observed

344 in VTE patients or in GnomAD achieved a CADD score above 15. This explains why the association could

345 not have been detected by the two other strategies based on fixed CADD score  $\geq 15$ . The median of

346 CADD scores observed for GnomAD variants in this region is 1.44 and the adjusted CADD scores of

347 selected variants range from 1.62 to 8.50. These observations emphasize the need to adapt thresholds

348 depending on the genomic region under analysis. Interestingly, only early-onset VTE patients carry

349 qualifying rare variants and have non-null WSS scores (Fig 3). Among early-onset patients, a trend is

350 also observed for WSS scores to decrease with increasing age of onset.

351 **Figure 3: WSS scores in the CADD region depending on the age at first VTE event.** The dashed line

352 corresponds to the age 50 discriminating early onset from late onset events.

353 The CADD region R126442 was then tested for association with 20 biological VTE biomarkers available

354 in MARTHA patients: antithrombin, basophil, eosinophil, Factor VIII, Factor XI, fibrinogen, hematocrit,

355 lymphocytes, mean corpuscular volume, mean platelet volume, monocytes, neutrophils, PAI-1,

356 platelets count, protein C, protein S, prothrombin time, red blood cells count, von Willebrand Factor,

357 and white blood cells count. For this, a linear regression model was used where adjustment was made

358 on age at sampling and sex. At the Bonferroni threshold of 0.0025, one significant association

359 ( $p = 7.1 \cdot 10^{-4}$ ) was observed, VTE patients with a non-null WSS score exhibiting decreased haematocrit

360 levels, a surrogate marker of red blood cells (Table in S3 Table). A similar trend ( $p = 4.6 \cdot 10^{-3}$ ) was

361 observed with red blood cell count.

362 We also investigated the association of the identified region with 376 plasma protein antibodies that

363 were selected to be involved in thrombosis-related processes and that have been previously profiled

364 in MARTHA (32,35). Regression analysis were conducted on log transformed values of antibodies and

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

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

389

## 390 Discussion

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

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

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

462 To our knowledge, CADD regions represent predefined testing units for RVAT that cover the highest  
463 proportion of the genome. These regions have been made publicly available (cf "Data availability"  
464 section below). CADD regions are part of a whole new strategy of rare variant analysis in the whole  
465 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/](https://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.

## 485 References

- 486 1. Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum  
487 statistic. *PLoS genetics*. 2009;5(2):e1000384.
- 488 2. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing  
489 data with the sequence kernel association test. *Am J Hum Genet*. 2011 Jul 15;89(1):82–93.
- 490 3. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and  
491 statistical tests. *Am J Hum Genet*. 2014 Jul 3;95(1):5–23.
- 492 4. Bellenguez C, Charbonnier C, Grenier-Boley B, Quenez O, Le Guennec K, Nicolas G, et al.  
493 Contribution to Alzheimer's disease risk of rare variants in TREM2, SORL1, and ABCA7 in 1779  
494 cases and 1273 controls. *Neurobiol Aging*. 2017 Nov;59:220.e1-220.e9.
- 495 5. Shaffer JR, LeClair J, Carlson JC, Feingold E, Buxó CJ, Christensen K, et al. Association of low-  
496 frequency genetic variants in regulatory regions with nonsyndromic orofacial clefts. *American  
497 Journal of Medical Genetics Part A*. 2019 Mar;179(3):467–74.
- 498 6. Wang Q, Dhindsa RS, Carss K, Harper AR, Nag A, Tachmazidou I, et al. Rare variant contribution  
499 to human disease in 281,104 UK Biobank exomes. *Nature* [Internet]. 2021 Aug 10 [cited 2021  
500 Aug 12]; Available from: <https://www.nature.com/articles/s41586-021-03855-y>
- 501 7. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases:  
502 application to analysis of sequence data. *Am J Hum Genet*. 2008 Sep;83(3):311–21.
- 503 8. Bis JC, Jian X, Kunkle BW, Chen Y, Hamilton-Nelson KL, Bush WS, et al. Whole exome sequencing  
504 study identifies novel rare and common Alzheimer's-Associated variants involved in immune  
505 response and transcriptional regulation. *Mol Psychiatry*. 2018 Aug 14;
- 506 9. Cirulli ET, White S, Read RW, Elhanan G, Metcalf WJ, Tanudjaja F, et al. Genome-wide rare  
507 variant analysis for thousands of phenotypes in over 70,000 exomes from two cohorts. *Nat  
508 Commun* [Internet]. 2020 Jan 28 [cited 2020 May 18];11. Available from:  
509 <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6987107/>
- 510 10. Guey S, Kraemer M, Hervé D, Ludwig T, Kossorotoff M, Bergametti F, et al. Rare RNF213  
511 variants in the C-terminal region encompassing the RING-finger domain are associated with  
512 moyamoya angiopathy in Caucasians. *Eur J Hum Genet*. 2017;25(8):995–1003.
- 513 11. Bocher O, Génin E. Rare variant association testing in the non-coding genome. *Hum Genet*  
514 [Internet]. 2020 Jun 4 [cited 2020 Jun 8]; Available from:  
515 <http://link.springer.com/10.1007/s00439-020-02190-y>
- 516 12. UK10K Consortium, Walter K, Min JL, Huang J, Crooks L, Memari Y, et al. The UK10K project  
517 identifies rare variants in health and disease. *Nature*. 2015 Oct 1;526(7571):82–90.
- 518 13. Zhang F, Lupski JR. Non-coding genetic variants in human disease. *Hum Mol Genet*. 2015 Oct  
519 15;24(R1):R102-110.
- 520 14. Castel SE, Cervera A, Mohammadi P, Aguet F, Reverter F, Wolman A, et al. Modified penetrance  
521 of coding variants by cis-regulatory variation contributes to disease risk. *Nature Genetics*. 2018  
522 Sep;50(9):1327–34.

- 523 15. Morrison AC, Huang Z, Yu B, Metcalf G, Liu X, Ballantyne C, et al. Practical Approaches for  
524 Whole-Genome Sequence Analysis of Heart- and Blood-Related Traits. *The American Journal of*  
525 *Human Genetics*. 2017 Feb;100(2):205–15.
- 526 16. Cochran JN, Geier EG, Bonham LW, Newberry JS, Amaral MD, Thompson ML, et al. Non-coding  
527 and Loss-of-Function Coding Variants in TET2 are Associated with Multiple Neurodegenerative  
528 Diseases. *Am J Hum Genet*. 2020 May 7;106(5):632–45.
- 529 17. Li Z, Li X, Liu Y, Shen J, Chen H, Zhou H, et al. Dynamic Scan Procedure for Detecting Rare-  
530 Variant Association Regions in Whole-Genome Sequencing Studies. *The American Journal of*  
531 *Human Genetics*. 2019 May;104(5):802–14.
- 532 18. He Z, Xu B, Buxbaum J, Ionita-Laza I. A genome-wide scan statistic framework for whole-  
533 genome sequence data analysis. *Nature Communications*. 2019 Jul 9;10(1):1–11.
- 534 19. Bocher O, Marenne G, Pierre AS, Ludwig TE, Guey S, Tournier-Lasserve E, et al. Rare variant  
535 association testing for multicategory phenotype. *Genetic Epidemiology*. 2019;43(6):646–56.
- 536 20. Bocher O, Marenne G, Tournier-Lasserve E, FREX Consortium, Génin E, Perdry H. Extension of  
537 SKAT to multi-category phenotypes through a geometrical interpretation. *Eur J Hum Genet*  
538 [Internet]. 2021 Jan 14 [cited 2021 Jan 15]; Available from:  
539 <http://www.nature.com/articles/s41431-020-00792-8>
- 540 21. Havrilla JM, Pedersen BS, Layer RM, Quinlan AR. A map of constrained coding regions in the  
541 human genome. *Nature Genetics*. 2019 Jan;51(1):88–95.
- 542 22. Genome Aggregation Database Consortium, Karczewski KJ, Francioli LC, Tiao G, Cummings BB,  
543 Alföldi J, et al. The mutational constraint spectrum quantified from variation in 141,456  
544 humans. *Nature*. 2020 May;581(7809):434–43.
- 545 23. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness  
546 of variants throughout the human genome. *Nucleic Acids Research*. 2019 Jan 8;47(D1):D886–  
547 94.
- 548 24. Nishizaki SS, Boyle AP. Mining the Unknown: Assigning Function to Noncoding Single Nucleotide  
549 Polymorphisms. *Trends in Genetics*. 2017 Jan;33(1):34–45.
- 550 25. Pujar S, O’Leary NA, Farrell CM, Loveland JE, Mudge JM, Wallin C, et al. Consensus coding  
551 sequence (CCDS) database: a standardized set of human and mouse protein-coding regions  
552 supported by expert curation. *Nucleic Acids Res*. 2018 04;46(D1):D221–8.
- 553 26. Barrett LW, Fletcher S, Wilton SD. Regulation of eukaryotic gene expression by the untranslated  
554 gene regions and other non-coding elements. *Cell Mol Life Sci*. 2012 Nov;69(21):3613–34.
- 555 27. Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shores N, Adrian J, et al. Expanded encyclopaedias  
556 of DNA elements in the human and mouse genomes. *Nature*. 2020 Jul;583(7818):699–710.
- 557 28. Itan Y, Shang L, Boisson B, Ciancanelli MJ, Markle JG, Martinez-Barricarte R, et al. The mutation  
558 significance cutoff: gene-level thresholds for variant predictions. *Nat Methods*. 2016  
559 Feb;13(2):109–10.

- 560 29. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving  
561 access to variant interpretations and supporting evidence. *Nucleic Acids Res.* 2018 Jan  
562 4;46(D1):D1062–7.
- 563 30. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature.*  
564 2015 Oct 1;526(7571):68–74.
- 565 31. Lindström S, Wang L, Smith EN, Gordon W, van Hylckama Vlieg A, de Andrade M, et al. Genomic  
566 and transcriptomic association studies identify 16 novel susceptibility loci for venous  
567 thromboembolism. *Blood.* 2019 Nov 7;134(19):1645–57.
- 568 32. Razzaq M, Iglesias MJ, Ibrahim-Kosta M, Goumidi L, Soukarieh O, Proust C, et al. An artificial  
569 neural network approach integrating plasma proteomics and genetic data identifies PLXNA4 as  
570 a new susceptibility locus for pulmonary embolism. *Sci Rep.* 2021 Jul 7;11(1):14015.
- 571 33. Germain M, Saut N, Grelische N, Dina C, Lambert J-C, Perret C, et al. Genetics of Venous  
572 Thrombosis: Insights from a New Genome Wide Association Study. *PLOS ONE.* 2011 Sep  
573 27;6(9):e25581.
- 574 34. Roupie A-L, Dossier A, Goulenok T, Perozziello A, Papo T, Sacre K. First venous  
575 thromboembolism in admitted patients younger than 50years old. *European Journal of Internal  
576 Medicine.* 2016 Oct 1;34:e18–20.
- 577 35. Razzaq M, Goumidi L, Iglesias M-J, Munsch G, Bruzelius M, Ibrahim-Kosta M, et al. Explainable  
578 Artificial Neural Network for Recurrent Venous Thromboembolism Based on Plasma  
579 Proteomics. In: Cinquemani E, Paulevé L, editors. *Computational Methods in Systems Biology.*  
580 Cham: Springer International Publishing; 2021. p. 108–21. (Lecture Notes in Computer Science).
- 581 36. Li J, Ji L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation  
582 matrix. *Heredity.* 2005 Sep;95(3):221–7.
- 583 37. Cokic VP, Bhattacharya B, Beleslin-Cokic BB, Noguchi CT, Puri RK, Schechter AN. JAK-STAT and  
584 AKT pathway-coupled genes in erythroid progenitor cells through ontogeny. *J Transl Med.* 2012  
585 Jun 7;10:116.
- 586 38. Weiss G, Ganz T, Goodnough LT. Anemia of inflammation. *Blood.* 2019 Jan 3;133(1):40–50.
- 587 39. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al.  
588 Comprehensive mapping of long range interactions reveals folding principles of the human  
589 genome. *Science.* 2009 Oct 9;326(5950):289–93.
- 590 40. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The Human Genome  
591 Browser at UCSC. *Genome Res.* 2002 Jan 6;12(6):996–1006.
- 592 41. Chen L, Xie X, Zhang X, Jia W, Jian J, Song C, et al. The expression, regulation and adhesion  
593 function of a novel CD molecule, CD226, on human endothelial cells. *Life Sci.* 2003 Sep  
594 19;73(18):2373–82.
- 595 42. Kojima H, Kanada H, Shimizu S, Kasama E, Shibuya K, Nakauchi H, et al. CD226 mediates platelet  
596 and megakaryocytic cell adhesion to vascular endothelial cells. *J Biol Chem.* 2003 Sep  
597 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, Bernstein 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

## 626 Supporting Information captions

627 **S1 Fig. Steps performed in RAVA-FIRST.**

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

630 **S3 Fig. Percentage of CADD regions ( $\geq 1\text{kbp}$ ) overlapping each of the three genomic categories.**

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

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

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

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

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

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

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

645

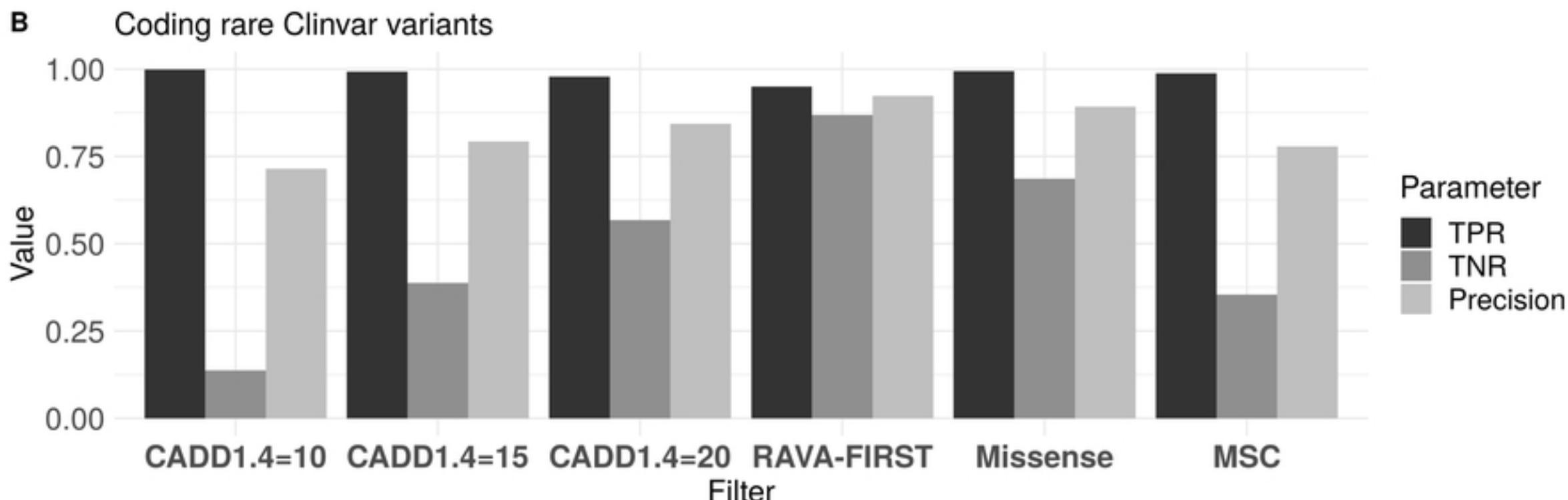
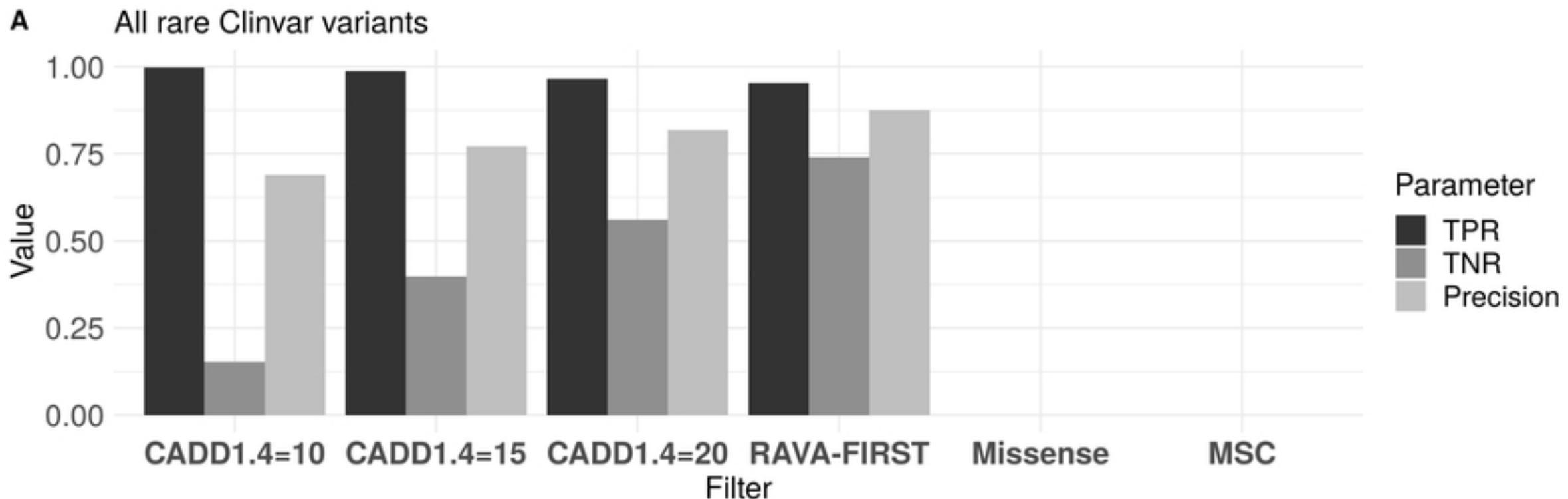


Figure 1

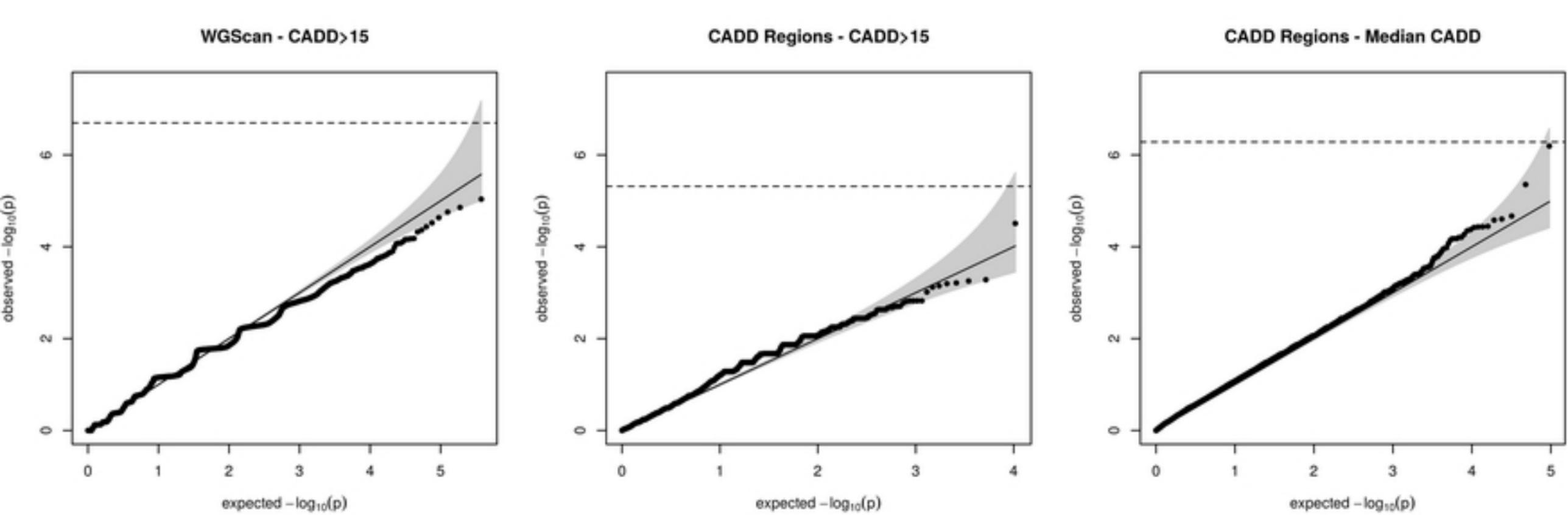


Figure 2

### WSS scores in CADD region 18:66788277-66809402

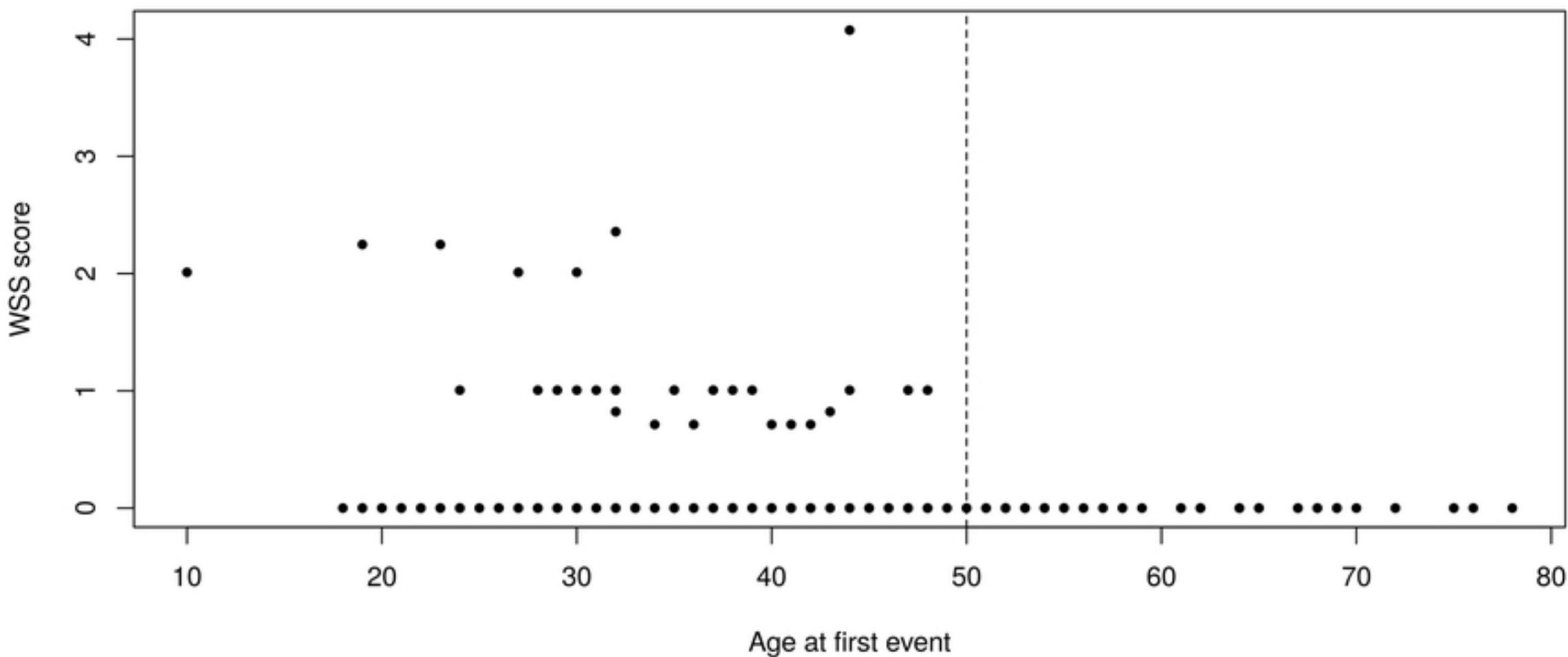


Figure 3