Whole genome sequencing identifies rare variants in *ANK1, LRRN1*, *HAS1*, and other genes and gene regulatory regions for stroke in type 1 diabetes

Anni A. Antikainen[1,2,3], Jani Haukka[1,2,3], Anmol Kumar[1,2,3], Anna Syreeni[1,2,3], Stefanie Hägg-Holmberg[1,2,3], Anni Ylinen[1,2,3], Elina Kilpeläinen[4], Anastasia Kytölä[4], Aarno Palotie[4,5,6], Jukka Putaala[7], Lena M. Thorn[1,2,3,8], Valma Harjutsalo[1,2,3], Per-Henrik Groop[1,2,3,9] and Niina Sandholm[1,2,3], on behalf of the FinnDiane Study Group

1. Folkhälsan Institute of Genetics, Folkhälsan Research Center, Helsinki, Finland

2. Department of Nephrology, University of Helsinki and Helsinki University Hospital, Helsinki, Finland

3. Research Program for Clinical and Molecular Metabolism, Faculty of Medicine, University of Helsinki, Helsinki, Finland

4. Institute for Molecular Medicine Finland (FIMM), HiLIFE, University of Helsinki, Helsinki, Finland

5. Analytic and Translational Genetics Unit, Department of Medicine, Department of Neurology and Department of Psychiatry, Massachusetts General Hospital, Boston, MA, USA

6.The Stanley Center for Psychiatric Research and Program in Medical and Population Genetics, The Broad Institute of MIT and Harvard, Cambridge, MA, USA

7. Neurology, Helsinki University Hospital and University of Helsinki, Helsinki, Finland

8. Department of General Practice and Primary Health Care, University of Helsinki and Helsinki University Hospital, Helsinki, Finland

9. Department of Diabetes, Central Clinical School, Monash University, Melbourne, Victoria, Australia

Corresponding Author:

Per-Henrik Groop, per-henrik.groop@helsinki.fi

Words: 5480

Tables: 3

Figures: 5

Abstract

Individuals with type 1 diabetes (T1D) carry a markedly increased risk of stroke, with distinct clinical and neuroimaging characteristics as compared to those without diabetes. Therefore, we studied stroke genetics in individuals with T1D using whole-genome sequencing (N=571) and whole-exome sequencing (N=480). We attempted replication of the lead discoveries in our T1D specific cohort using a genome-wide association study (N=3,945) and genotyping (N=3,600), and in the general population from the large-scale population-wide FinnGen project and UK Biobank summary statistics. We analysed the genome comprehensively with single-variant analyses, gene aggregate analyses, and aggregate analyses on genomic windows, enhancers and promoters. We discovered T1D specific stroke loci on 4q33-34.1, *SREBF1*, and *ANK1*; and stroke loci that likely generalize to the non-diabetic population on *HAS1*, *LRRN1*, *UACA*, *LTB4R*, *LINC01500* (promoter capture loop to *DACT1*), and *TRPM2-AS* promoter. We further validated the promoter with an *in vitro* cell-based assay.

Introduction

Stroke is a notable cause of mortality and long-term disability worldwide, with diabetes among the most important risk factors. Specifically, the standardized incidence ratio is nearly 6-fold among individuals with diabetes compared to the general population[1,2]. Furthermore, 537 million adults live with diabetes today and the incidence is still rising (IDF Diabetes Atlas, 2021)[3], highlighting the importance of diabetes as a risk factor for stroke. Even though much of this trend is driven by an increase in obesity and insulin-resistant type 2 diabetes (T2D), the incidence of insulin-dependent type 1 diabetes (T1D) has increased as well[4,5]. T1D is a lifelong condition caused by an autoimmune reaction towards the pancreas and treated with daily insulin injections. Interestingly, differences regarding the stroke risk imposed by diabetes have been observed between the two diabetes types. The strokes themselves may be of hemorrhagic (20%) or ischemic (80%) origin and classified into even more specific subtypes. T1D increases the risk of both ischemic- and hemorrhagic stroke[6,7], while the risk imposed by T2D has been estimated more modest for hemorrhagic strokes[7,8]. Importantly, T1D predisposes individuals particularly towards cerebral small-vessel disease and strokes of microvascular origin[9,10]. Diabetes causes also other complications, of which diabetic kidney disease (DKD) and severe retinopathy predict cerebrovascular disease in T1D[11]. Understanding stroke pathophysiology in diabetes is important for improving treatment and quality of life for individuals with T1D.

Stroke heritability has been estimated to vary between 30% and 40% in the general population[12]. To date, 126 common genomic loci have been associated with stroke with genome-wide significance, although partly lacking external replication: 65 common genomic loci had been associated with stroke, a stroke subtype, or small vessel disease[13], while recently, a large cross-ancestry genome-wide association study (GWAS) meta-analysis proposed 61 novel genomic loci for stroke[14]. Stroke heritability varies greatly depending on the subtype, with the largest heritability estimated for large artery atherosclerotic stroke and the lowest for small vessel disease[15]. Associations at many of the known common stroke loci overlap with other cardiovascular phenotypes, e.g., coronary artery disease (CAD)[12]. In addition, some of the stroke-

37 associated genes have been linked to diabetic complications, e.g., *COL4A1* to DKD[16,17], or *SH2B3* to T1D
38 itself[12,18]. So far, very few studies have investigated genetic risk factors for stroke in diabetes[19–21], and no
39 genome-wide studies on stroke exist in individuals with diabetes, although we previously speculated that
40 there may be a heritable component as a history of maternal stroke was associated with hemorrhagic stroke
41 in individuals with T1D[22]. On the other hand, genetic studies on CAD in diabetes have identified a few
42 diabetes-specific loci[23,24], although still pending external replication, and have replicated three known
43 general population CAD risk loci in diabetes: *CDKN2B-AS1*, *PSRC1* and *LPA*[21,23,25].

44 A substantial proportion of heritability remains unexplained for stroke[12,15]. Importantly, stroke is a complex
45 phenotype with multiple subtypes, which complicates genetic studies and reduces statistical power.
46 Furthermore, some rare monogenic disorders have stroke as one of their manifestations[12,13]. In fact, rare
47 genetic variants with minor allele frequency (MAF) of ≤1% may significantly contribute to stroke
48 heritability. In GWASs, the imputation accuracy of rare variants may be limited, and largely depends on
49 the minor allele count (MAC) in the reference sample[26]. Rare variants can be reliably studied with next-
50 generation sequencing-based techniques such as whole-genome sequencing (WGS) and whole-exome
51 sequencing (WES). We have previously used WES to study low frequency variants for DKD[27], and
52 identified protein coding variants associated in lipid and apolipoprotein traits in T1D[28]. In the general
53 population, novel stroke risk loci have been identified with WGS[29], although Jurgens et al. (2022) analyzed
54 WES in the UK Biobank for cardiometabolic traits and did not discover exome-wide significant stroke risk
55 genes[30].

56 Here we studied stroke genetics with WGS and WES in individuals with T1D with multiple statistical
57 approaches by focusing on rare and low-frequency genomic variants. We performed single variant and
58 aggregate analyses both within genes and on intergenic regions. Gene aggregate analyses were performed
59 separately for predicted protein truncating- and protein altering variants (PTV and PAV, respectively;
60 omnibus test, SKAT-O), while genomic windows, enhancers and promoters were analyzed with a
61 functional annotation weighted sliding-window approach (STAAR-O). We performed the analyses in

62 Finnish individuals with T1D and well-characterized stroke phenotypes. Historically, the Finnish
63 population has been isolated and, thus, represents a unique genetic background with enrichment of low-
64 frequency deleterious variants[31], which may in part enable the discovery of rare disease-associated
65 mutations. We aimed both to find stroke-risk loci specific to individuals with T1D, and to identify risk loci
66 generalizable to the non-diabetic population, since discovery of rare variants is more probable in a high-
67 risk Finnish diabetic population. Finally, we performed cell-based *in vitro* experiments to further validate
68 a discovered promoter region. Altogether, here we report the first genome-wide study on stroke genetics in
69 diabetes.

70

Results

Single variant analyses

This study is part of the Finnish Diabetic Nephropathy (FinnDiane) Study[32]. We studied WGS in 571 and WES in 480 non-related individuals with T1D, entailing 112 and 74 stroke cases, respectively (**Table 1, Supplementary Fig. 1, 2**). We sought for variants associated with stroke using fixed-effects score test meta-analysis of WES and WGS data for variants available in both data sets, and Firth regression on variants available only in the other (MAC≥5, **Fig. 1**). We discovered a suggestive stroke-risk locus, 4q33-34.1, with minimally adjusted additive inheritance model (4:170787127, *p*-value=8.83×10^{-8}, MAF=3.7%, **Table 2**, **Fig. 2**). We attempted replication in individuals with T1D in the FinnDiane GWAS data set (N=3,945), restricted to high imputation quality variants (r^2>0.80), where we had good statistical power (>80%) to detect a nominal association with an odds ratio (OR) ≥2.5 for additive low-frequency variants (MAF=1%) (**Extended Data Fig. 1**). However, the variant 4:170787127 was unavailable for replication in the T1D specific GWAS and in the FinnGen general population GWAS summary statistics (**Supplementary Note**). A variant with the third lowest *p*-value on 4q33-34.1 did not replicate in T1D nor the general population (**Table 2**).

As DKD is a common diabetic complication that has been reported to predict incident stroke in T1D[11], we performed additional analyses adjusted for DKD, and discovered a rare missense mutation on *SREBF1* exome-wide significantly (*p*-value<3×10^{-7}) associated with stroke (rs114001633, p.Pro227Leu, *p*-value=7.30×10^{-8}, MAF=0.26%) (**Extended Data Fig. 2, Table 2**). In the stroke subtype analysis for the lead findings, this variant was genome-wide significant for hemorrhagic stroke (*p*-value=8.96×10^{-9}, MAC=3, **Supplementary Table 1**). However, rs114001633 did not pass the MAC threshold in the stroke subtype analysis (MAC≥5), thus, the result must be interpreted with caution. Due to the rarity of the variant, we performed additional genotyping for replication, whereby the variant replicated for hemorrhagic stroke (*p*-value=0.02, N=3,263).

95 Recessive variants can have high mutation severity in comparison to e.g., autosomal dominant variants[33].
96 Therefore, we further considered a recessive inheritance model and identified a genome-wide significant
97 variant on *DHX8* intron with the minimal adjustment (rs1728177, *p*-value=7.19×10^{-9}, MAF=26%)
98 (**Extended Data Fig. 3**, **Table 2**), which however did not replicate. Furthermore, we found two suggestive
99 recessive loci, including one with the minimal adjustment on 5' untranslated region (UTR) of *LTB4R*
100 (rs2224123, *p*-value=4.77×10^{-7}, MAF=14%). rs2224123 did not replicate in T1D but was borderline
101 significant in the general population with an additive model (FinnGen GWAS: *p*-value=0.052). In the
102 sequencing data, significance with an additive model was nominal (*p*-value=0.019), suggesting that
103 rs2224123 acts recessively, but recessive data were not available for the replication look-up in the general
104 population.

105 **Gene aggregate analyses**

106 In order to improve statistical power for rare (MAF≤1%) and low-frequency (MAF≤5%) variants, we
107 performed gene aggregate analyses with an optimal unified sequence kernel association test (SKAT-O) and
108 aimed to identify combined exonic variant effects associated with stroke in two distinct classes: PAV and
109 PTV i.e. the more severe putative loss-of-function variants[34] (**Supplementary Note**). We required a
110 cumulative MAC (CMAC) ≥5 across the included variants in each gene. In the minimally adjusted model,
111 low-frequency PAVs on *ANK1* were significantly associated with stroke (*p*-value=2.23×10^{-6}, CMAC=247),
112 even more strongly with ischemic stroke (*p*-value=1.31×10^{-6}, CMAC=225) (**Fig. 3A, Extended Data Fig.
113 4, Supplementary Tables 2 and 3**). Furthermore, the aggregate of low-frequency or rare PAVs was
114 suggestively associated with stroke in nine genes (**Fig. 3A**). Of these, *TARBP2* associated significantly with
115 ischemic stroke (*p*-value=1.71×10^{-7}, CMAC=5, MAF≤1%), and *CLEC4M* with hemorrhagic stroke (*p*-
116 value=4.74×10^{-15}, CMAC=11, MAF≤1%).

117 After additional DKD adjustment, rare PAVs in *LRRN1* were significantly associated with stroke (*p*-
118 value=3.49×10^{-6}, CMAC=15), more strongly with ischemic stroke (*p*-value=8.69×10^{-6}, CMAC=12; **Fig.**

119 **3B**, **Extended Data Fig. 5, Supplementary Tables 2 and 4**). Furthermore, we identified suggestive genes,
120 missed in the minimally adjusted model: *MAP3K12* and *MTRNR2L7*. In the stroke subtype analysis, rare
121 PAVs in *MAP3K12* were significantly associated with ischemic stroke (p-value=1.72×10^{-7}, CMAC=17),
122 and in *MTRNR2L7* with hemorrhagic stroke (p-value=2.24×10^{-6}, CMAC=6). *MAP3K12* and *TARBP2* are
123 located close to each other on the genome, thus, they may represent the same association signal through
124 linkage disequilibrium (LD) or modifier effects onto the causal gene (**Supplementary Fig. 3**).

125 The aggregate of PTVs was suggestively associated with stroke in two genes including hyaluronan synthase
126 1 (*HAS1;* **Fig. 3A**, **Supplementary Table 3**). In analysis for stroke subtypes, *HAS1* associated significantly
127 with ischemic stroke (p-value=7.39×10^{-7}, CMAC=7). With additional DKD adjustment, rare PTVs in *HAS1*
128 (p-value=3.11×10^{-5}) and *UACA* (p-value=6.77×10^{-5}, CMAC=6), and low-frequency PTVs in *ARPC5* (p-
129 value=4.15×10^{-5}, CMAC=39), were significantly associated with stroke (**Fig. 3B, Supplementary Table**
130 **4**).

131 **Replication of gene aggregate findings**

132 Most variants within the lead genes are rare or ultra-rare (MAF≈0.1%), making replication with imputed
133 genomic data problematic. Nevertheless, we attempted replication of the significant gene aggregate
134 discoveries within the FinnDiane GWAS data by including also five directly genotyped rare variants;
135 replication analyses were performed with SKAT-O. We also attempted replication for the variants
136 individually but had only limited power to detect an association even with nominal significance (p-
137 value<0.05) for variants with MAF=0.1% and OR<9 (**Extended Data Fig. 1**). Nevertheless, *ANK1* and
138 *LRRN1* showcased weak evidence of replication in T1D. Although *ANK1* did not reach significance with
139 SKAT-O (**Supplementary Table 5**), one of the available fifteen variants was significant for stroke
140 (rs779805849, p-value=0.01), and two additional variants replicated for hemorrhagic stroke (rs146416859
141 and rs61753679, p-value<0.05) (**Table 3, Fig. 4, Supplementary Table 3**). *LRRN1* did not replicate in
142 FinnDiane with rare PAVs (p-value=0.50, $N_{variant}$=4) (**Supplementary Table 6**). However, when we

extended the model to low-frequency PAVs (**Supplementary Table 6**), thus improved statistical power and imputation quality, *LRRN1* replicated for ischemic stroke after DKD adjustment (*p*-value=0.039, $N_{variant}$=6) in line with the sequencing analyses indicating an association with ischemic stroke. *UACA* contained two rare PTVs associated with stroke, of which one replicated through genotyping (*p*-value=0.0030, **Supplemental Table 4**). However, the variant was ultra-rare (MAC=1), and replication thus uncertain. *ARPC5* did not replicate in T1D. In addition, we were unable to replicate *HAS1* in T1D due to missing data; we directly genotyped one variant but found no rare allele carriers.

We further attempted replication in the general population by look-ups from two UK Biobank general population WES studies[30,35] (**Supplementary Tables 7 and 8**). Importantly, *HAS1* replicated for stroke in both studies with ultra-rare loss-of-function variants (MAF≤0.1%: Jurgens et al. *p*-value=0.039[30]; Backman et al. *p*-value=0.035[35]), while *UACA* replicated only in the latter study (MAF≤0.01%: Backman et al. *p*-value=0.035[35]). Finally, *LRRN1* replicated for stroke with an ultra-rare missense variant model (MAF≤0.001%: Backman et al. *p*-value=0.026[35]), although not for ischemic stroke. Of note, *ANK1* did not replicate in the general population.

Out of the suggestive genes, *FOXO1*, *TARBP2*, and *MAP3K12* showcased weak replication in T1D (FinnDiane GWAS) (**Supplementary Tables 3, 4 and 5**). One variant within *FOXO1* replicated for hemorrhagic stroke (*p*-value=0.012), two within *MAP3K12* for hemorrhagic stroke (*p*-value=0.013); and *TARBP2* replicated for hemorrhagic stroke with SKAT-O (*p*-value=2.59×10^{-4}). UK Biobank general population gene burden WES analysis look-ups supported stroke associations for *UTS2*, *MAP3K12*, and *FOXO1* (**Supplementary Table 8**)[30,35], while one variant within *RRP9* replicated for stroke in the FinnGen summary statistics.

Known Mendelian stroke genes in T1D

Mutations on Mendelian stroke risk genes may for instance cause small vessel disease or cerebral cavernous malformations, which can eventually lead to stroke[12]. We inspected the association of 17 autosomal genes

167 previously linked to stroke through nonsynonymous mutations[36], and found 3 to also associate with stroke
168 in T1D (p-value<0.05; **Extended Data Fig. 6**). Rare PAVs on *KRIT1* associated with stroke (p-
169 value=0.018), particularly with ischemic stroke (p-value=0.0092). Furthermore, rare PAVs on *ADA2* and
170 on *TREX1* associated with hemorrhagic stroke (p-value=0.027 and p-value=0.010, respectively). Loss-of-
171 function mutations on *KRIT1* cause vascular malformations, while *ADA2* has been linked to
172 autoinflammatory small vessel vasculitis and *TREX1* to small vessel disease[12,36].

173 **Sliding window analyses**

174 In GWASs, roughly 88% of the identified disease-associated variants are located within non-coding
175 regions[37]. To increase statistical power to detect low-frequency and rare variants on non-coding regulatory
176 regions, we performed genome-wide sliding-window aggregate analyses within 4,000 base pair (bp)
177 regions, by statistically up-weighting functionally important variants according to Combined Annotation-
178 Dependent Depletion (CADD) functional annotations and rarity with STAAR-O[38,39] (**Extended Data Fig.**
179 **7**). We found further evidence for the 4q33-34.1 genomic region as we discovered fourteen windows, within
180 the extended region, with a genome-wide significant association between an aggregate of low-frequency
181 variants and stroke (**Supplementary Table 9, Fig. 5A**). Importantly, two of these windows (4:170782001-
182 170786000, p-value=3.40×10^{-8}, CMAC=934; and 4:170784001-170786000, p-value=1.10×10^{-8},
183 CMAC=1190) and ten individual variants within the 4q33-34.1 genomic region replicated for stroke in T1D
184 (FinnDiane GWAS: p-value<0.05; **Supplementary Table 10**). To identify the most likely effector genes
185 for the 4q33-34.1, we inspected variant expression quantitative trait loci (eQTL) from GTEx Portal and
186 eQTLGen Consortium[40], and functional genomics from 3D Genome Browser[41]. 4q33-34.1 is located in the
187 same topologically associating domain with distal promoters of *GALNTL6*, *MFAP3L* and *AADAT* in the
188 frontal lobe and hippocampus (**Supplementary Fig. 4**). In addition, promoter capture high-throughput
189 chromosome conformation capture (PCHi-C) links could be identified for a few individual variants, e.g.,
190 for *GALNTL6* in the hippocampus, and *AADAT* and *MFAP3L* in the dorsolateral prefrontal cortex. Several

191 variants were eQTLs for *LINC02431* in testis. In addition, one variant was a significant eQTL of *AADAT*
192 in esophagus (normalized effect size [NES] =-0.55); and a nominal eQTL in many other tissues.

193 When we inspected rare variants (MAF≤1%), we discovered multiple suggestive windows, e.g., close to or
194 within the *CNTN1*, *CNTN4*, *LINC01500*, and *TGOLN2* genes (**Supplementary Table 9**, **Fig. 5B**). In stroke
195 subtype analysis, the *CNTN1* window was genome-wide significant for hemorrhagic stroke (12:40950001-
196 40954000: *p*-value=2.10×10^{-8}, CMAC=24). Interestingly, *CNTN1* and *CNTN4* are located on different
197 chromosomes, but belong to the same protein family encoding contactins[42]; however, replication is still
198 pending. Importantly, a suggestive window near *LINC01500* (14:59004001-59008000: *p*-value=2.53×10^{-7},
199 CMAC=19) replicated for stroke in T1D (FinnDiane GWAS: *p*-value=0.01, CMAC=24). Four variants
200 within the window were available in the FinnGen GWAS, and one replicated (rs1281241634, *p*-value=0.03)
201 (**Supplementary Table 10**). According to PCHi-C, the *LINC01500* intronic window looped to the *DACT1*
202 promoter on the dorsolateral prefrontal cortex (**Supplementary Fig. 5**). Finally, the *TGOLN2* window
203 replicated for hemorrhagic stroke in T1D (FinnDiane GWAS: *p*-value=0.037).

204 **Promoters and enhancers**

205 As a more targeted approach to explore the non-coding regulatory genome, we studied rare and low-
206 frequency variants in established regulatory regions – enhancers and promoters – as defined in FANTOM5
207 cap analysis of gene expression (CAGE) human raw data[43–45], with promoters defined as the transcription
208 start site (TSS) extended to 1,000 bp. We discovered three enhancers with suggestive stroke-associated
209 enrichment of rare or low-frequency variants within intronic regions of *TRPM3*, *LOC105378983,* and
210 *BDNF*, encoding brain-derived neurotrophic factor (**Supplementary Tables 11 and 12, Extended Data**
211 **Fig. 8**). Regional aggregate replications were not possible in the T1D specific GWAS ($N_{variant}$<2). In
212 addition, individual variants were missing or did not replicate. PCHi-C linked the *BDNF* enhancer to its
213 promoter on the dorsolateral prefrontal cortex and hippocampus (**Supplementary Fig. 5**). The *BDNF*
214 enhancer was significant after multiple testing correction for ischemic stroke (*p*-value=1.01×10^{-6}).

215 We did not identify stroke-associated promoters after multiple testing correction (p-value<3×10^{-7},
216 **Extended Data Fig. 9**). The strongest associations were two *TGOLN2* promoters (p-value=5.60×10^{-6},
217 CMAC=9, MAF≤1%), located on the previously mentioned *TGOLN2* window, and a *TRPM2-AS* promoter
218 (p-value=5.78×10^{-6}, CMAC=33, MAF≤1%; **Supplementary Tables 13 and 14**). *TGOLN2* promoters were
219 borderline significant for ischemic stroke (p-value=4.71×10^{-7}, MAC=5), but did not replicate in T1D.
220 *TRPM2-AS* promoter nearly replicated in T1D (FinnDiane GWAS: p-value=0.053). When we inspected
221 variants individually, one out of nine available variants replicated in the general population for ischemic
222 stroke (FinnGen GWAS: p-value=0.038). In GTEx, rs762428 within the *TRPM2-AS* promoter associated
223 significantly to *TRPM2* level in whole blood (NES=-0.63) and lungs (NES=-0.41, p<0.001), also nominally
224 in other tissues such as the hypothalamus (NES=-0.42). *TRPM2* encodes a calcium-permeable and non-
225 selective cation channel expressed mainly in the brain[46], which has been linked to ischemic stroke and
226 belongs to the same melastatin subfamily as the previously mentioned *TRPM3*[46,47]. *TRPM2* inhibitors have
227 been proposed as a drug target option for central nervous system diseases[48], thus, our results suggested that
228 these inhibitors could be beneficial also for stroke in T1D.

229 Our cell-based experimental research on the *TRPM2-AS* promoter detected *TRPM2-AS* and *TRPM2*
230 transcripts specifically in HELA cells among tested cell-lines (**Extended Data Fig. 10**). Luciferase
231 promoter analysis of the *TRPM2-AS* promoter in HELA cells indicated strong promoter activity; although,
232 the most strongly stroke-associated variant, rs753589764, did not significantly affect luciferase activity
233 under normal cell culture conditions (p-value=0.27, 22 technical repeats). However, we cannot rule out a
234 variant effect under cellular stress, e.g., oxidative stress; a hallmark of cardiovascular disease in T1D[49].

235 # Discussion

236 Stroke heritability has been estimated to range between 30% and 40%, but the genomic loci identified thus
237 far explain only a small fraction of heritability[12]. One potential explanation underlying the missing
238 heritability are rare variants missed by GWAS. Therefore, we performed WES and WGS in a total of 1,051
239 Finnish individuals with T1D to discover rare and low-frequency variants associated with stroke and its
240 major subtypes, either specific for T1D, or generalizable to the non-diabetic population. We identified
241 multiple significant loci with evidence of replication, including protein altering or truncating variants on
242 *ANK1*, *HAS1*, *UACA*, and *LRRN1*, as well as a 4q33-34.1 intergenic region.

243 With single variant analyses, we identified a suggestively stroke-associated missense mutation on *SREBF1*
244 (rs114001633, p.Pro227Leu) with an additive inheritance model adjusted for DKD, which was genome-
245 wide significant for hemorrhagic stroke, and further replicated for hemorrhagic stroke. Variant penetrance
246 was 8% in replication, i.e., 1 out of 12 individuals carrying the risk allele developed hemorrhagic stroke.
247 As the variant was ultra-rare, and we had a relatively small number of hemorrhagic stroke cases, further
248 replication is needed in T1D to conform this finding. *SREBF1* encodes a transcription factor involved in
249 lipid metabolism and insulin signaling[50]. In addition, we found a suggestively stroke-associated recessive
250 variant near *LTB4R*, encoding the leukotriene B4 receptor, which nearly replicated in the general population
251 with an additive inheritance model. Importantly, recessive replication is still pending.

252 Gene aggregate tests (SKAT-O) detected four genes with significant stroke-associated burden of PAVs
253 (*ANK1* and *LRRN1*) or PTVs (*HAS1* and *UACA*) and evidence of replication; *LRRN1*, *HAS1*, and *UACA*
254 after adjustment for DKD. *ANK1* did not replicate in T1D with SKAT-O, however, one out of the fifteen
255 available variants replicated for stroke in T1D (rs779805849, p.Val136Glu). Of note, SIFT and PolyPhen
256 predicted many *ANK1* variants as deleterious[51,52]. *ANK1* encodes ankyrin-1, within which mutations cause
257 hereditary spherocytosis[53]. Previous genome-wide association studies have linked the gene to T2D[54], while
258 another gene from the ankyrin protein family, *ANK2*, is a previously identified stroke risk locus[55]. Rare

259 PAVs in *LRRN1* associated with ischemic stroke. *LRRN1* did not replicate with the corresponding model
260 in T1D, however; with a model extended to low-frequency PAVs, *LRRN1* replicated for ischemic stroke.
261 Rare variant replication is problematic with GWAS data due to the uncertainty of the imputation, which
262 may explain the need of increasing the allele frequency threshold to observe a successful replication.
263 Furthermore, *LRRN1* replicated with ultra-rare variants in the general population (UK Biobank WES:
264 MAF≤0.001%, *p*-value=0.025)[35]. *LRRN1* encodes leucine rich repeat neuronal protein 1, with a brain-
265 enriched expression profile. *HAS1* consistently replicated in the general population (UK Biobank WES:
266 MAF≤0.1%, *p*-value=0.012; MAF≤1%, *p*-value=0.035)[30,35], while *UACA* replicated in one study with ultra-
267 rare variants (UK Biobank WES: MAF≤0.01%, *p*-value=0.035)[35]. *HAS1* encodes an enzyme producing
268 hyaluronan and with expression induced by inflammation and glycemic stress[56]. Of note, an increased
269 hyaluronan turnover has been suggested to follow ischemic stroke[57]. *HAS1* replication was not feasible in
270 T1D due to absent mutation carriers, thus, a diabetes-specific replication is pending. Nevertheless, *HAS1*
271 PTVs may be of particular importance in T1D; dysregulation of endothelial glycocalyx hyaluronan has
272 been suggested to contribute to diabetic complications[58]. Finally, it must be noted that PTVs have not been
273 functionally confirmed as loss-of-function, but the annotations are predictions; PTV at the beginning of a
274 gene is likely more severe than at the end, and in fact, PTVs nearer to the *HAS1* transcription start site were
275 more strongly associated with stroke.

276 It has been suggested that disease-associated single nucleotide variants are more likely to occur on
277 intergenic regulatory regions than on exons[37]. To increase statistical power on regulatory regions, we
278 performed statistical aggregate tests in genomic windows, enhancers and promoters[38,45]. Of note, we
279 extended genomic window length from the default to increase statistical power, which however also
280 reduced precision as the causal region might be narrower. We found fourteen genome-wide significant
281 stroke-associated windows with low-frequency variants on 4q33-34.1, of which two replicated for stroke
282 in T1D. According to eQTLs and PCHi-C interactions, 4q33-34.1 variants most likely target *GALNTL6*,
283 *MFAP3L* or *AADAT*, which encodes α-aminoadipate aminotransferase. We also discovered a suggestive

284 stroke-associated window within *LINC01500*, which replicated for stroke in T1D. According to PCHi-C,
285 the *LINC01500* window targets a promoter of *DACT1*, encoding a Wnt signaling repressor[59]. Finally, we
286 identified a suggestive stroke-associated promoter of *TRPM2-AS*, which nearly replicated in T1D (*p*-
287 value=0.053). Importantly, transient receptor melastatin 2 (*TRPM2*) has been previously associated with
288 ischemic stroke[48]. Our functional cell-based assay validated the promoter activity of the *TRPM2-AS* region.
289 However, the most strongly stroke-associated variant, rs753589764, did not associate with *TRPM2-AS*
290 promoter activity in normal cell culture conditions. Interestingly, rare variants on an enhancer region within
291 *TRPM3* gene – from the same protein family but on a different chromosome – associated suggestively with
292 stroke (p=1.34×10^{-5}).

293 Limitations of the study include the limited statistical power at the discovery stage, especially for the stroke
294 subtypes, and replication of rare variants with imputed GWAS data. We were able to improve the statistical
295 power on exomes by meta-analyzing WES and WGS, and we performed stroke-subtype specific analyses
296 only for a limited number of suggestive findings to avoid spurious signals due to unstable statistical
297 estimates. To further improve statistical power, we performed statistical aggregate tests on gene exons and
298 on intergenic regions, i.e., enhancers, promoters, and genomic windows. Of note, we studied only
299 transcribed enhancers, and thus, some enhancers could have been missed. We defined promoters with an
300 arbitrarily selected 1,000 bp extension downstream TSS, which may not have always been optimal as the
301 promoter lengths vary. A further limitation is the lack of sequencing-based replication data in individuals
302 with T1D. Instead, we sought for replication by combining available data sources, i.e., FinnGen (Finnish
303 general population GWAS), UK Biobank (general population WES), and FinnDiane (GWAS and
304 genotyping in Finnish individuals with T1D).

305 The strengths of this study include a well characterized cohort and comprehensively performed single
306 variant and aggregate analyses both for the coding and non-coding regions of the genome. Stroke is a
307 challenging phenotype to address with ICD codes and many loci associated with rare stroke phenotypes
308 may go unnoticed even with large population-wide genetic studies. We performed analyses for well-defined

309 stroke phenotypes verified by trained neurologists. Furthermore, as we conducted the analyses in specific
310 high-risk individuals from an isolated population, thus with lesser genetic and phenotypic diversity, we had
311 improved statistical opportunities to identify genetic risk loci.

312 In conclusion, we studied rare and low-frequency stroke-associated variants with WES and WGS in
313 individuals with T1D and report the first genome-wide study on stroke genetics in diabetes. The results
314 highlight 4q33-34.1, *SREBF1*, and *ANK1* for stroke in T1D; and *HAS1*, *UACA*, *LRRN1*, *LINC01500*, and
315 *TRPM2-AS* promoter as stroke risk loci that likely generalize to the non-diabetic population.

316 # Methods

317 **Participants and stroke definitions**

318 The study is part of the FinnDiane Study established to investigate complications of T1D. We have access
319 to 490 and 583 non-related individuals with T1D with WES and WGS, respectively. Individuals in the
320 present study were diagnosed with T1D by their attending physician and had diabetes onset age <40 and
321 insulin initiated within one calendar year from the diabetes diagnosis. Stroke phenotypes were identified
322 from the Finnish Death Registry, Statistics Finland, and the Care Register for Health Care, Finnish Institute
323 for Health and Welfare until the end of 2017 and were verified by trained neurologists from medical files
324 and brain imaging data. We required acute stroke events to have occurred after the T1D diagnosis. For
325 individuals without data verified by neurologists available (N_{WGS}=27, N_{WES}=2), we considered only the
326 registry data, and excluded five controls with transient ischemic attacks or other intermediate stroke
327 phenotypes due to potential for misclassification (**Supplementary Note**). Furthermore, we classified stroke
328 events into ischemic- and hemorrhagic strokes whenever stroke subtypes verified by neurologists were
329 available. Ischemic strokes entailed lacunar-, non-lacunar-, and unclear infarctions, while hemorrhagic
330 strokes comprised intracerebral and subarachnoid hemorrhages. We limited controls to individuals with
331 >35 years of age and >20 years of diabetes duration, because only ~5% of cases had experienced their first
332 stroke event before these limits in sequencing data (**Table 1, Supplementary Table 15**).

333 **Sequencing material**

334 In total, 599 individuals were whole genome sequenced at Macrogen Inc. using the Illumina HiSeq X
335 platform (Macrogen Inc., Rockville, MD, USA) with at least 30× average coverage (1-8 lanes).
336 Furthermore, 502 participants were whole-exome sequenced at the University of Oxford, UK, as described
337 earlier[27]. In short, libraries were multiplexed and captured with Illumina TruSeq™ Exome Enrichment Kit
338 and sequenced on an Illumina HiSeq2000 with 100 bp paired end reads (1-2 lanes). Seven samples did not
339 pass the initial quality control (QC) or showed discrepancies with other genetic data. For this work, the pre-

340 processed aligned reads of 495 individuals were converted back to unaligned FASTQ-reads and re-
341 processed with comparable pipeline to WGS.

342 First, we trimmed the WES and WGS sequencing reads[60]. We processed 495 WES and 599 WGS samples
343 further according to Broad Institute's best practices guidelines with Genome Analysis Toolkit 4 (GATK4)[61]
344 (**Supplementary Fig. 6**): We aligned reads by lane to GRCh38 reference genome with Burrows-Wheeler
345 Aligner, sorted and marked duplicate reads, recalibrated bases by chromosome, and called variants by
346 sample. WES and WGS were joint called separately. We filtered variants according to excess
347 heterozygosity threshold of 54.69, truth sensitivity level 99.7%, and GATK's recommended tranche
348 thresholds for SNPs and indels. We assessed sample concordance with FinnDiane GWAS data whenever
349 possible. Finally, 490 and 583 individuals passed QC within WES and WGS, respectively. In variant QC,
350 for autosomal variants, we required Hardy-Weinberg equilibrium (HWE) *p*-value $>10^{-10}$ and variant call
351 rate >98%; and for X chromosome variants, only variant call rate >98%. We annotated variants with
352 SNPEff v.5 software[62]. WES and WGS comprised 324,817 and 21.92 million variants, respectively.

353 **Replication material**

354 In FinnDiane, we have GWAS data for 6,458 Finnish individuals with T1D and their relatives. Genomes
355 were genotyped at the University of Virginia, and previously processed to GRCh37 reference genome[63].
356 We shifted the genotyping positions to GRCh38, re-imputed the data to SISu v3 reference panel, and
357 annotated with SNPEff v.5 software[62] (**Supplementary Fig. 7**). We restricted the data to 15,026,805 high
358 imputation quality variants (r^2>0.80), and individuals to those with T1D; age at onset <40 years and insulin
359 treatment initiated within two years from diagnosis, if insulin initiation year known, and to those not in the
360 sequencing data. Controls were required to have diabetes duration >20 years and age >35 years. GWAS
361 replication entailed 367 cases and 3,578 controls (**Supplementary Fig. 8, Supplementary Tables 16 and**
362 **17**). Instead of principal component adjustment, we utilized rvtests kinship matrix with Balding-Nichol's
363 approximation in single variant analysis and GEMMA relatedness matrix in aggregate analyses[64–66].

364 Twelve lead variants were selected for replication by variant genotyping in 3,600 FinnDiane participants
365 with T1D on one Agena iPlex multiplexing assay at the Institute for Molecular Medicine Finland, Helsinki,
366 Finland (**Supplementary Table 18**). Variants were prioritized onto to the same multiplexing design based
367 on statistical significance, and one heterozygous variant carrier within sequencing data included as the
368 positive control. Replication was limited to individuals within GWAS data to perform relatedness
369 adjustment with kinship matrix (N=3,263, **Supplementary Table 19, Supplementary Fig. 9**).

370 # Statistical methods

371 **Single variant analyses**

372 We analyzed variants available in WES and WGS data with score test fixed-effect inverse variance based
373 meta-analysis (CMAC≥5, WES and WGS: MAC≥2), and variants only available in one of the data sets with
374 Firth regression (MAC≥5). Furthermore, autosomal variants were analyzed recessively according to similar
375 scheme (cumulative homozygote minor allele carriers ≥5; WES and WGS homozygote carriers ≥2).
376 Analyses were carried out with rvtests (version 20190205)[65] and metal (version 20110325)[67], except
377 autosomal recessive Firth regression with plink2 (version 20210420)[68,69].

378 **Gene aggregate analyses**

379 We performed autosomal gene aggregate tests with SKAT-O, combining burden and variance-component
380 tests, with the aim to increase statistical power and stability[70]. We performed SKAT-O meta-analysis,
381 separately with PAVs and PTVs, between WES and WGS using MetaSKAT (version 0.81)[71], which
382 exploits single variant score statistics. PTVs were here defined as predicted loss-of-function mutations
383 (**Supplementary Note**). PAV analyses entail, in addition to the PTVs, mutations that are predicted to alter
384 amino acid sequence. Only variable sites (MAC≥1) were accepted into gene aggregate ($N_{variant}$≥2,
385 CMAC≥5). We did not report genes with all variants in perfect LD, and inspected individual variant stroke-
386 associations within the genes using score test fixed-effects meta-analysis[65,67]. Multiple testing correction

387 based on number of included genes resulted in significance thresholds of p-value<4×10^{-6} for PAVs
388 (MAF≤1% and MAF≤5%), p-value<7×10^{-5} for PTVs with MAF≤1%, and p-value<5×10^{-5} for PTVs with
389 MAF≤5%. For the known Mendelian stroke risk genes[36], we reported results regardless of variant number
390 or CMAC.

391 **Sliding-window analyses and regulatory regions**

392 We performed functional annotation weighted sliding window analyses on the WGS data with the STAAR
393 R package 0.9.6[38]. For variant annotations we utilized CADD v1.6 GRCh38 data[39,72], more specifically;
394 variant MAF (to up-weight rarer variants), pre-computed CADD score, and the first annotation principal
395 components from seven annotation classes (**Extended Data Fig. 7, Supplementary Note**). Following
396 guidelines[38], missing annotations were imputed to default and distributions were standardized before
397 principal component analysis. Annotations were transformed to PHRED scale with appropriate direction.
398 We utilized 4,000 bp windows ($N_{variant}$≥2, CMAC≥5) separated by 2,000 bp skips. We studied 20.7 million
399 autosomal variants with available functional annotations.

400 As enhancers and promoters, we considered FANTOM5 CAGE profiles reprocessed to the GRCh38
401 reference genome (https://fantom.gsc.riken.jp/5/)[43–45]; however, we extended CAGE TSSs to form full-
402 length promoters (1,000 bp). We analyzed regulatory regions ($N_{variant}$≥2, CMAC≥5) with the STAAR R
403 package[38], using only allele frequencies as variant annotations, which allowed us to include more autosomal
404 variants. With low-frequency variants, 184,192 promoters and 24,045 enhancers were analyzed, resulting
405 in multiple testing corrected significance thresholds p-value<2.9×10^{-7} and p-value<2.6×10^{-6}, respectively.
406 For rare variants, the thresholds were p-value<3.5×10^{-7} and p-value<4.3×10^{-6}, respectively. We did not
407 report regions with all variants in perfect LD.

408 **Regional plots and functional characterization**

409 We produced regional association plots with LocusZoom[73] and Gviz R package 1.38.3[74]; and inspected
410 variant characteristics from GTEx Portal, eQTLGen Consortium (p-value<0.05)[40], RegulomeDB[75], YUE
411 Lab (http://3dgenome.fsm.northwestern.edu/)[41], and the Ensemble Variant Effect Predictor[51,52,76].

412 **Functional research on *TRPM2-AS* promoter**

413 HELA, HEK-293 and HUVEC cells were cultured as per standard protocol. Total RNA was extracted with
414 TRIzol™ Reagent, cDNA synthesized with SuperScript™ III Reverse Transcriptase, and qPCR performed
415 with iTaq™ Universal SYBR® Green Supermix as per manufacturer's protocol (**Supplementary Note**).
416 *TRPM2-AS* promoter sequence was PCR amplified from genomic DNA of one FinnDiane participant
417 carrying rs753589764 minor allele, and one participant carrying none of the identified *TRPM2-AS* rare
418 minor alleles, and cloned on pBV-Luc plasmid (Addgene plasmid #16539). Transfection of *TRPM2-AS*
419 promoter sequence carrying pBV-Luc reporter and renilla luciferase control plasmid was done using
420 FuGENE® HD Transfection Reagent: Four technical repeats each transfection.

421 **Single variant replication**

422 We performed GWAS in individuals with T1D using score test (rvtests version)[65], and calculated statistical
423 power with the genpwr R package 1.0.4[77]. Genotyping was successful for a varying number of individuals
424 depending on the variant. We observed mutation carriers for six variants (**Supplementary Table 18**) and
425 performed single variant analyses similarly with score test. We analyzed one *LRRN1* variant with linear
426 regression (stats R package 4.2.1) and without relatedness adjustment, because the only mutation carrier
427 was missing in the GWAS data i.e., kinship matrix.

428 We attempted general population replication from the FinnGen project release 6
429 (https://www.finngen.fi/en), and selected stroke phenotypes that best matched our definitions
430 (**Supplementary Note**).

431 **Gene aggregate replication**

432 We performed SKAT-O analyses in FinnDiane GWAS (r^2>0.80) for T1D specific replication with the
433 GMMAT R package 1.3.2[78], by including also genotyped variants, and imputed missing data to mean (i.e.,
434 individuals with missing genotype at a variant). We attempted general population replication from UK
435 Biobank WES studies[30,35].

436 **Sliding-window and regulatory region replication**

437 We attempted T1D specific replication with GWAS data (r^2>0.80) using STAAR R package[38].

Data availability

Individual-level data for the study participants are not publicly available, because of the restrictions due to the study consent provided by the participant at the time of data collection. The readers may propose collaboration to research the individual level data with correspondence with the lead investigator. Gene aggregate test stroke summary statistics will be provided in the Supplementary Data.

Code availability

We utilized public softwares for the statistical analyses: plink2 (https://www.cog-genomics.org/plink/2.0/), metal (http://csg.sph.umich.edu/abecasis/metal/), rvtests (http://zhanxw.github.io/rvtests/), metaSKAT (https://cran.r-project.org/web/packages/MetaSKAT/), STAAR-O (https://github.com/xihaoli/STAAR), and GMMAT (https://github.com/hanchenphd/GMMAT).

References

1. Harjutsalo, V. *et al.* Presence and Determinants of Cardiovascular Disease and Mortality in Individuals With Type 1 Diabetes of Long Duration: The FinnDiane 50 Years of Diabetes Study. *Diabetes Care* **44**, 1885–1893 (2021).
2. Giorda, C. B. *et al.* Incidence and Risk Factors for Stroke in Type 2 Diabetic Patients. *Stroke* **38**, 1154–1160 (2007).
3. Sun, H. *et al.* IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes Res. Clin. Pract.* **183**, (2022).
4. Forlenza, G. P. & Rewers, M. The epidemic of type 1 diabetes: what is it telling us? *Curr. Opin. Endocrinol. Diabetes Obes.* **18**, 248–251 (2011).
5. Harjutsalo, V., Sund, R., Knip, M. & Groop, P.-H. Incidence of type 1 diabetes in Finland. *Jama* **310**, 427–428 (2013).
6. Ståhl, C. H. *et al.* Glycaemic control and excess risk of ischaemic and haemorrhagic stroke in patients with type 1 diabetes: a cohort study of 33 453 patients. *J. Intern. Med.* **281**, 261–272 (2017).
7. Janghorbani, M. *et al.* Prospective study of type 1 and type 2 diabetes and risk of stroke subtypes: the Nurses' Health Study. *Diabetes Care* **30**, 1730–1735 (2007).
8. Muñoz-Rivas, N. *et al.* Type 2 Diabetes and Hemorrhagic Stroke: A Population-Based Study in Spain from 2003 to 2012. *J. Stroke Cerebrovasc. Dis.* **25**, 1431–1443 (2016).
9. Thorn, L. M. *et al.* Clinical and MRI Features of Cerebral Small-Vessel Disease in Type 1 Diabetes. *Diabetes Care* **42**, 327–330 (2019).
10. Putaala, J. *et al.* Diabetes mellitus and ischemic stroke in the young. *Neurology* **76**, 1831–1837 (2011).

469 11. Hägg, S. *et al.* Incidence of stroke according to presence of diabetic nephropathy and severe
470 diabetic retinopathy in patients with type 1 diabetes. *Diabetes Care* **36**, 4140–4146 (2013).
471 12. Dichgans, M., Pulit, S. L. & Rosand, J. Stroke genetics: discovery, biology, and clinical
472 applications. *Lancet Neurol.* **18**, 587–599 (2019).
473 13. Debette, S. & Markus, H. S. Stroke Genetics: Discovery, Insight Into Mechanisms, and Clinical
474 Perspectives. *Circ. Res.* **130**, 1095–1111 (2022).
475 14. Mishra, A. *et al.* Stroke genetics informs drug discovery and risk prediction across ancestries.
476 *Nature* (2022) doi:10.1038/s41586-022-05165-3.
477 15. Malik, R. & Dichgans, M. Challenges and opportunities in stroke genetics. *Cardiovasc. Res.* **114**,
478 1226–1240 (2018).
479 16. Lanfranconi, S. & Markus, H. S. COL4A1 mutations as a monogenic cause of cerebral small vessel
480 disease: a systematic review. *Stroke* **41**, e513–e518 (2010).
481 17. Ewens, K. G., George, R. A., Sharma, K., Ziyadeh, F. N. & Spielman, R. S. Assessment of 115
482 candidate genes for diabetic nephropathy by transmission/disequilibrium test. *Diabetes* **54**, 3305–
483 3318 (2005).
484 18. Todd, J. A. *et al.* Robust associations of four new chromosome regions from genome-wide
485 analyses of type 1 diabetes. *Nat. Genet.* **39**, 857–864 (2007).
486 19. Syreeni, A. *et al.* Haptoglobin Genotype Does Not Confer a Risk of Stroke in Type 1 Diabetes.
487 *Diabetes* db220327 (2022) doi:10.2337/db22-0327.
488 20. Dahlström, E. H. *et al.* The Low-Expression Variant of FABP4 Is Associated With Cardiovascular
489 Disease in Type 1 Diabetes. *Diabetes* **70**, 2391–2401 (2021).
490 21. Vujkovic, M. *et al.* Discovery of 318 new risk loci for type 2 diabetes and related vascular
491 outcomes among 1.4 million participants in a multi-ancestry meta-analysis. *Nat. Genet.* **52**, 680–691
492 (2020).

493 22. Ylinen, A. *et al.* The impact of parental risk factors on the risk of stroke in type 1 diabetes. *Acta*
494 *Diabetol.* **58**, 911–917 (2021).
495 23. Antikainen, A. A. V. *et al.* Genome-wide association study on coronary artery disease in type 1
496 diabetes suggests beta-defensin 127 as a risk locus. *Cardiovasc. Res.* **117**, 600–612 (2021).
497 24. Qi, L. *et al.* Association Between a Genetic Variant Related to Glutamic Acid Metabolism and
498 Coronary Heart Disease in Individuals With Type 2 Diabetes. *JAMA* **310**, 821–828 (2013).
499 25. Fall, T., Gustafsson, S., Orho-Melander, M. & Ingelsson, E. Genome-wide association study of
500 coronary artery disease among individuals with diabetes: the UK Biobank. *Diabetologia* **61**, 2174–
501 2179 (2018).
502 26. Si, Y., Vanderwerff, B. & Zöllner, S. Why are rare variants hard to impute? Coalescent models
503 reveal theoretical limits in existing algorithms. *Genetics* **217**, iyab011 (2021).
504 27. Sandholm, N. *et al.* The Genetic Landscape of Renal Complications in Type 1 Diabetes. *J. Am.*
505 *Soc. Nephrol.* **28**, 557–574 (2017).
506 28. Sandholm, N. *et al.* Whole-exome sequencing identifies novel protein-altering variants
507 associated with serum apolipoprotein and lipid concentrations. *Genome Med.* **14**, 132 (2022).
508 29. Hu, Y. *et al.* Whole-Genome Sequencing Association Analyses of Stroke and Its Subtypes in
509 Ancestrally Diverse Populations From Trans-Omics for Precision Medicine Project. *Stroke* **53**, 875–885
510 (2022).
511 30. Jurgens, S. J. *et al.* Analysis of rare genetic variation underlying cardiometabolic diseases and
512 traits among 200,000 individuals in the UK Biobank. *Nat. Genet.* **54**, 240–250 (2022).
513 31. Locke, A. E. *et al.* Exome sequencing of Finnish isolates enhances rare-variant association power.
514 *Nature* **572**, 323–328 (2019).
515 32. Thorn, L. M. *et al.* Metabolic Syndrome in Type 1 Diabetes: Association with diabetic
516 nephropathy and glycemic control (the FinnDiane study). *Diabetes Care* **28**, 2019–2024 (2005).

517 33. Pei, J., Kinch, L. N., Otwinowski, Z. & Grishin, N. V. Mutation severity spectrum of rare alleles in
518 the human genome is predictive of disease type. *PLoS Comput. Biol.* **16**, e1007775 (2020).
519 34. Rivas, M. A. *et al.* Effect of predicted protein-truncating genetic variants on the human
520 transcriptome. *Science* **348**, 666–669 (2015).
521 35. Backman, J. D. *et al.* Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature*
522 **599**, 628–634 (2021).
523 36. Grami, N. *et al.* Global assessment of Mendelian stroke genetic prevalence in 101 635 individuals
524 from 7 ethnic groups. *Stroke* **51**, 1290–1293 (2020).
525 37. Edwards, S. L., Beesley, J., French, J. D. & Dunning, A. M. Beyond GWASs: illuminating the dark
526 road from association to function. *Am. J. Hum. Genet.* **93**, 779–797 (2013).
527 38. Li, X. *et al.* Dynamic incorporation of multiple in silico functional annotations empowers rare
528 variant association analysis of large whole-genome sequencing studies at scale. *Nat. Genet.* **52**, 969–
529 983 (2020).
530 39. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the
531 deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2019).
532 40. Võsa, U. *et al.* Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and
533 polygenic scores that regulate blood gene expression. *Nat. Genet.* **53**, 1300–1310 (2021).
534 41. Wang, Y. *et al.* The 3D Genome Browser: a web-based browser for visualizing 3D genome
535 organization and long-range chromatin interactions. *Genome Biol.* **19**, 151 (2018).
536 42. Morris, J. *et al.* Exploring the Role of Contactins across Psychological, Psychiatric and
537 Cardiometabolic Traits within UK Biobank. *Genes* **11**, (2020).
538 43. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature*
539 **507**, 455–461 (2014).

540 44. The FANTOM Consortium and the RIKEN PMI and CLST (DGT). A promoter-level mammalian
541 expression atlas. *Nature* **507**, 462–470 (2014).
542 45. Abugessaisa, I. *et al.* FANTOM5 CAGE profiles of human and mouse reprocessed for GRCh38 and
543 GRCm38 genome assemblies. *Sci. Data* **4**, 1–10 (2017).
544 46. Fleig, A. & Penner, R. The TRPM ion channel subfamily: molecular, biophysical and functional
545 features. *Trends Pharmacol. Sci.* **25**, 633–639 (2004).
546 47. Zong, P., Lin, Q., Feng, J. & Yue, L. A Systemic Review of the Integral Role of TRPM2 in Ischemic
547 Stroke: From Upstream Risk Factors to Ultimate Neuronal Death. *Cells* **11**, 491 (2022).
548 48. Belrose, J. C. & Jackson, M. F. TRPM2: a candidate therapeutic target for treating neurological
549 diseases. *Acta Pharmacol. Sin.* **39**, 722–732 (2018).
550 49. Tang, W. H. W. *et al.* Oxidative Stress and Cardiovascular Risk in Type 1 Diabetes Mellitus:
551 Insights From the DCCT/EDIC Study. *J. Am. Heart Assoc.* **7**, e008368 (2018).
552 50. DeBose-Boyd, R. A. & Ye, J. SREBPs in Lipid Metabolism, Insulin Signaling, and Beyond. *Trends*
553 *Biochem. Sci.* **43**, 358–368 (2018).
554 51. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat.*
555 *Methods* **7**, 248–249 (2010).
556 52. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on
557 protein function using the SIFT algorithm. *Nat. Protoc.* **4**, 1073–1081 (2009).
558 53. Park, J. *et al.* Mutational characteristics of ANK1 and SPTB genes in hereditary spherocytosis.
559 *Clin. Genet.* **90**, 69–78 (2016).
560 54. Yan, R. *et al.* A novel type 2 diabetes risk allele increases the promoter activity of the muscle-
561 specific small ankyrin 1 gene. *Sci. Rep.* **6**, 25105 (2016).
562 55. Malik, R. *et al.* Multiancestry genome-wide association study of 520,000 subjects identifies 32
563 loci associated with stroke and stroke subtypes. *Nat. Genet.* **50**, 524–537 (2018).

564 56. Siiskonen, H., Oikari, S., Pasonen-Seppänen, S. & Rilla, K. Hyaluronan Synthase 1: A Mysterious
565 Enzyme with Unexpected Functions. *Front. Immunol.* **6**, 1–11 (2015).
566 57. Katarzyna Greda, A. & Nowicka, D. Hyaluronidase inhibition accelerates functional recovery
567 from stroke in the mouse brain. *J. Neurochem.* **157**, 781–801 (2021).
568 58. Wang, G., Tiemeier, G. L., van den Berg, B. M. & Rabelink, T. J. Endothelial Glycocalyx
569 Hyaluronan: Regulation and Role in Prevention of Diabetic Complications. *Am. J. Pathol.* **190**, 781–
570 790 (2020).
571 59. Zhang, L., Gao, X., Wen, J., Ning, Y. & Chen, Y.-G. Dapper 1 Antagonizes Wnt Signaling by
572 Promoting Dishevelled Degradation*. *J. Biol. Chem.* **281**, 8607–8612 (2006).
573 60. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence
574 data. *Bioinformatics* **30**, 2114–2120 (2014).
575 61. Van der Auwera, G. & O'Connor, B. *Genomics in the Cloud: Using Docker, GATK, and WDL in*
576 *Terra*. (O-Reilly Media, 2020).
577 62. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide
578 polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3.
579 *Fly (Austin)* **6**, 80–92 (2012).
580 63. Salem, R. M. *et al.* Genome-Wide Association Study of Diabetic Kidney Disease Highlights Biology
581 Involved in Glomerular Basement Membrane Collagen. *J. Am. Soc. Nephrol.* **30**, 2000 (2019).
582 64. Balding, D. J. & Nichols, R. A. A method for quantifying differentiation between populations at
583 multi-allelic loci and its implications for investigating identity and paternity. *Genetica* **96**, 3–12 (1995).
584 65. Zhan, X., Hu, Y., Li, B., Abecasis, G. R. & Liu, D. J. RVTESTS: an efficient and comprehensive tool
585 for rare variant association analysis using sequence data. *Bioinformatics* **32**, 1423–1426 (2016).
586 66. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies.
587 *Nat. Genet.* **44**, 821–824 (2012).

588 67. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide
589 association scans. *Bioinformatics* **26**, 2190–2191 (2010).
590 68. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets.
591 *Gigascience* **4**, s13742-015 (2015).
592 69. Purcell, S. & Chang, C. PLINK v2.00a3LM.
593 70. Lee, S. *et al.* Optimal unified approach for rare-variant association testing with application to
594 small-sample case-control whole-exome sequencing studies. *Am. J. Hum. Genet.* **91**, 224–237 (2012).
595 71. Lee, S., Teslovich, T. M., Boehnke, M. & Lin, X. General framework for meta-analysis of rare
596 variants in sequencing association studies. *Am. J. Hum. Genet.* **93**, 42–53 (2013).
597 72. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human
598 genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
599 73. Pruim, R. J. *et al.* LocusZoom: regional visualization of genome-wide association scan results.
600 *Bioinformatics* **26**, 2336–2337 (2010).
601 74. Hahne, F. & Ivanek, R. Visualizing genomic data using Gviz and bioconductor. in *Statistical*
602 *genomics* 335–351 (Springer, 2016).
603 75. Boyle, A. P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB.
604 *Genome Res.* **22**, 1790–1797 (2012).
605 76. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
606 77. Moore, C. M., Jacobson, S. A. & Fingerlin, T. E. Power and Sample Size Calculations for Genetic
607 Association Studies in the Presence of Genetic Model Misspecification. *Hum. Hered.* **84**, 256–271
608 (2019).
609 78. Chen, H. *et al.* Efficient Variant Set Mixed Model Association Tests for Continuous and Binary
610 Traits in Large-Scale Whole-Genome Sequencing Studies. *Am. J. Hum. Genet.* **104**, 260–274 (2019).

Acknowledgements

We are indebted to the late Carol Forsblom (1964–2022), the international coordinator of the FinnDiane Study Group, for his considerable contribution. The skilled technical assistance of Maikki Parkkonen, Heli Krigsman, Mira Rahkonen, Anna Sandelin, and Jaana Tuomikangas is gratefully acknowledged. We also want to acknowledge all the physicians and nurses at each FinnDiane center throughout Finland participating in the recruitment and characterization of the individuals with T1D (**Supplementary Table 20**) as well as all the FinnDiane participants. In addition, we acknowledge the participants and investigators of the FinnGen study. We acknowledge that the ELIXIR Finland node, hosted at the CSC – IT Center for Science for ICT resources, enabled the WES and WGS data processing. Finally, we want to acknowledge Bert Vogelstein and Jukka Kallijärvi for material provided for the *in vitro* promoter experiments: pBV-Luc plasmids (Bert Vogelstein) and renilla control plasmid (Kallijärvi lab, Folkhälsan Research Center).

We utilized data provided by GTEx. The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data reported here were obtained from the GTEx Portal on 01/04/2022: https://gtexportal.org/home/.

This work was supported by grants from Folkhälsan Research Foundation; Wilhelm and Else Stockmann Foundation; "Liv och Hälsa" Society; Sigrid Juselius Foundation; Helsinki University Central Hospital Research Funds [TYH2018207]; Novo Nordisk Foundation [NNF OC0013659], Academy of Finland [299200 and 316664]; European Foundation for the Study of Diabetes (EFSD) Young Investigator Research Award funds; an EFSD award supported by EFSD/Sanofi European Diabetes Research Programme in Macrovascular Complications; Finnish Foundation for Cardiovascular Research; Sigrid Juselius foundation (220027); and the Finnish Diabetes Research Foundation.

Author information

Affiliations and contributions.

Ethics declaration

Ethical approval

The study protocol has been approved by the ethics committee of the Helsinki and Uusimaa Hospital District (491/E5/2006, 238/13/03/00/2015, and HUS-3313-2018), and performed in accordance with the Declaration of Helsinki. All participants gave informed consent before participation.

Competing interests

P-H G has received investigator-initiated research grants from Eli Lilly and Roche, is an advisory board member for AbbVie, Astellas, AstraZeneca, Bayer, Boehringer Ingelheim, Cebix, Eli Lilly, Janssen, Medscape, Merck Sharp & Dohme, Mundipharma, Nestlé, Novartis, Novo Nordisk and Sanofi; and has received lecture fees from AstraZeneca, Bayer, Boehringer Ingelheim, Eli Lilly, Elo Water, Genzyme, Merck Sharp & Dohme, Medscape, Novartis, Novo Nordisk, PeerVoice, Sanofi, and Sciarc. Other authors declare no competing interests.

Tables

Table 1. Clinical characteristics of study participants in the next-generation sequencing data sets.

	WES			WGS		
	Cases	**Controls**	***p*-value**	**Cases**	**Controls**	***p*-value**
N	74	406		112	459	
Hemorrhagic/Ischemic	22/49			26/64		
CVD death <2017 (yes/no, % yes)	35/39 (47%)	66/340 (16%)	2.92×10^{-8}	58/54 (52%)	83/376 (18%)	2.61×10^{-12}
Sex (male/female, % males)	39/35 (53%)	182/224 (45%)	0.25	80/32 (71%)	236/223 (51%)	0.00013
Age (years)	48.39 (11.89)	60.69 (11.13)	7.08×10^{-13}	51.08 (10.29)	58.70 (9.61)	3.52×10^{-11}
T1D duration (years)	33.6 (10.45)	47.49 (9.84)	5.89×10^{-18}	36.16 (9.81)	46.00 (8.14)	6.34×10^{-18}
Calendar year of T1D onset*	1969.5 (11.75)	1968 (11)	0.29	1968 (10.25)	1969 (9)	0.80
T1D onset age (years)	14.79 (6.64)	13.21 (7.27)	0.066	14.92 (8.93)	12.69 (7.75)	0.016
Weighted mean HbA_{1c} (%)*	9.00 (2)	8.51 (1.4)	0.0037	8.86 (1.76)	8.34 (1.57)	0.00097
HbA_{1c} count*	19 (28.5)	29 (26)	0.0038	19 (30)	29 (30)	0.011
DKD (yes/no, yes-%)	52/22 (70%)	201/205 (50%)	0.00098	87/25 (78%)	206/253 (45%)	2.49×10^{-10}

*Weighted mean HbA1c is calculated until the stroke event or the end of follow-up. DKD = end-stage renal disease (ESRD), macro- or microalbuminuria. Mean (standard deviation, SD), *Median (interquartile range, IQR). Student's t-test, Wilcoxon signed rank test or Fisher's exact test.*

Table 2. Lead variants discovered with single variant association analyses.

SNP ID	Position	Potential target gene	Data	REF /ALT	MAF	N_{Hom}	OR [95% CI]	I^2	*p*-value	GWAS OR (*p*-value)	FinnGen OR (*p*-value)
Additive model	**(minimal adjustment)**										
rs376936219	4:170784011	*AADAT, MFAP3L, GALNTL6*	WGS	G/T	4.6%	-	4.77 [2.66-8.60]	-	1.67×10^{-7}	0.88 (0.35)	1.03 (0.34)
rs4435704	4:170787127		WGS	T/C	3.7%	-	6.04 [3.12-11.67]	-	8.83×10^{-8}	-	-
rs4401420	4:170787143		WGS	C/T	4.1%	-	5.16 [2.79-9.55]	-	1.76×10^{-7}	-	-
Additive model	**(DKD adjustment)**										
rs114001633	17:17819659	*SREBF1* (p.Pro227Leu)	++	G/A	0.3%	-	3804.35 [189.15-76515.45]	35%	7.30×10^{-8}	1.41 (0.30)	1.03 (0.87)
Recessive model	**(minimal adjustment)**										
rs2224123	14:24313752	*LTB4R**	++	A/T	14.1%	26	16.75 [5.59-50.17]	77%	4.77×10^{-7}	0.97 (0.92)	1.03 (0.052)
rs1728177	17:43605161	*DHX8, MEOX1, ETV4**	WGS	G/A	26.3%	46	6.94 [3.60-13.37]	-	7.19×10^{-9}	1.03 (0.87)	0.99 (0.59)
Recessive model	**(DKD adjustment)**										
rs76073237	1:160870594	*ITLN1, CD244*	WGS	GTA/G	36.7%	66	4.69 [2.58-8.54]	-	4.00×10^{-7}	-	-

*Multiple variants reported from a locus if the lead variant was unavailable in GWAS replication (in such case, we report the closest variants). If the variant was discovered already in the minimal model, we do not report DKD adjusted results. Variants are in Hardy-Weinberg equilibrium (p-value>0.05). N_{Hom}=Number of recessive carriers, ++Meta-analysis with positive minor allele effect direction, *Multiple eQTLs (**Supplemental Table 18**) or dorsolateral prefrontal cortex PCHi-C links. REF=Reference allele, ALT=Alternative allele, I^2=Meta-analysis heterogeneity estimate, OR=Odds ratio, CI=confidence interval.*

Table 3. Protein altering- and protein truncating variants in significant stroke risk genes according to significance thresholds corrected by the number of genes within designated models.

				Sequencing			FinnDiane GWAS		FinnGen GWAS	
Gene (Model)	**SNP ID**	**Consequence (SIFT/PolyPhen)**	**REF/ALT**	**MAC**	**OR [95% CI]**	***p*-value**	**MAC**	**OR (*p*-value)**	**MAC**	**OR (*p*-value)**
HAS1 (Minimal: PTV ≤1%)	19:51713664	p.Tyr500* (-/-)	G/T	-/2	0.31 [0.0050-18.88]	0.57				
	19:51713853	p.Cys437* (-/-)	G/T	-/1	0.32 [0.00062-162.99]	0.72				
	rs1396240967	Splice donor (-/-)	C/T	-/1	901.45 [3.53-230240.57]	0.016				
	rs58597876	Splice donor (-/-)	C/CACAC ACACACA	3/-	369.81 [17.90-7640.36]	0.00013				
LRRN1 (Minimal + DKD: PAV≤1%)	rs755350940	5 prime UTR (-/-)	C/T	2/1	4.45 [0.17-114.63]	0.37	20	1.05 (0.93)	1483	0.98 (0.85)
	rs142381203	p.Arg185Cys (0.00/0.79)	C/T	1/-	67.29 [0.67-6787.51]	0.074	1	0.82 (0.46**)		
	rs147007969	p.Thr190Ile (0.18/0.54)	C/T	2/4	33.55 [3.92-287.48]	0.0014	25	0.75 (0.61)	1431	0.94 (0.57)
	rs141825989	p.Arg305Cys (0.01/0.80)	C/T	-/1	439.22 [2.19-88150.42]	0.025	1	0.52 (0.81*)		
	rs770349026	p.Arg390His (0.00/0.98)	G/A	-/2	28.22 [0.44-1817.11]	0.12				
	rs747455683	p.Thr442Met (0.12/0.65)	C/T	1/-	2351.95 [6.73-821987.12]	0.0094			281	0.75 (0.35)
	rs543878366	p.Glu481Lys (0.01/0.27)	G/A	-/1	0.32 [0.00049-216.02]	0.74				
ANK1 (Minimal: PAV≤5%)	rs750580242	p.Ser896Arg (0.40/0.11)	A/T	1/-	24.02 [0.35-1656.73]	0.14				
	rs751965455	p.Met1607Ile (0.06/0.07)	C/T	-/1	0.29 [0.0017-49.02]	0.64				
	rs146416859	p.Arg1577Cys (0.03/0.01)	G/A	1/1	6.18 [0.20-194.73]	0.30	15	1.66 (0.49)	716	1.01 (0.94)
	rs34664882	p.Ala114Val (0.04/0.74)	G/A	13/15	3.16 [1.14-8.76]	0.028	141	1.12 (0.63)	8250	0.93 (0.17)
	rs199760447	p.Val1450Ile (1.00/0.00)	C/T	-/1	5229.13 [11.67-2344071.66]	0.0060	6	0.26 (0.24)	257	1.12 (0.70)
	rs201439151	p.Tyr1386His (0.01/0.50)	A/G	2/1	0.28 [0.016-4.63]	0.37	11	0.73 (0.71)	917	0.76 (0.062)
	rs10093583	p.Met159Val (0.40/0.00)	T/C	4/2	2.62 [0.33-20.88]	0.36	41	0.80 (0.61)	2757	1.03 (0.72)
	rs779805849	p.Val136Glu (0.00/0.99)	A/T	3/-	3.48 [0.15-78.91]	0.43	5	24.19 (0.017)	684	0.85 (0.34)
	rs148275567	p.Arg89Cys (0.00/1.00)	G/A	3/4	8.50 [1.36-53.12]	0.022	11	0.98 (0.98)	1056	0.85 (0.23)
	rs117516263	Structural interaction variant (-/-)	G/A	19/12	0.87 [0.35-2.17]	0.77	114	1.39 (0.22)	6980	0.92 (0.11)
	rs202226361	p.Val1048Met (0.00/1.00)	C/T	5/-	0.84 [0.099-7.09]	0.87	21	3.30 (0.055)	1510	1.01 (0.97)
	rs767717861	p.Thr1022Met (0.00/0.93)	G/A	-/1	11695.97 [20.42-6698538.99]	0.0038				
	8:41696385	p.Pro302Thr (0.00/1.00)	G/T	-/3	0.30 [0.012-7.50]	0.46				
	8:41696393	p.Ala299Glu (1.00/0.91)	G/T	-/3	6.74 [0.21-211.38]	0.28				
	rs2304877	p.Arg619His (0.01/0.003)	C/T	32/41	1.35 [0.72-2.54]	0.35	317	1.00 (0.98)	20065	1.00 (0.91)
	rs61753679	Structural interaction variant (-/-)	G/T	20/22	0.92 [0.40-2.08]	0.84	120	1.21 (0.47)	8343	0.99 (0.79)
	rs34387324	Structural interaction variant (-/-)	G/A	1/2	0.30 [0.012-7.35]	0.46	10	1.10 (0.92)	1206	1.19 (0.18)
	rs142690258	p.Asn528Ser (0.84/0.001)	T/C	7/5	4.25 [0.99-18.25]	0.051	21	0.69 (0.55)	2327	0.97 (0.72)
	rs140085544	p.Val463Ile (0.04/0.38)	C/T	1/2	143.88 [4.63-4468.82]	0.0046	17	2.43 (0.20)	1101	0.93 (0.60)
	rs769619019	p.Met476Leu (0.11/0.99)	T/A	-/1	0.30 [0.0015-58.17]	0.65				
	rs1015648649	p.Ala461Val (0.00/1.00)	G/A	1/-	0.31 [0.0010-94.56]	0.69				
	8:41717683	p.His442Arg (0.00/1.00)	T/C	1/-	0.31 [0.00099-99.04]	0.69				
	rs1364511169	p.His416Arg (1.00/1.00)	T/C	1/-	0.27 [0.0025-29.56]	0.59				
	rs61735313	p.Asn251Lys (0.01/1.00)	G/T	11/4	4.11 [1.08-15.61]	0.038	77	0.82 (0.53)	3998	0.99 (0.85)
UACA (Minimal + DKD: PTV≤1%)	rs781623644	p.Gln1116* (-/-)	G/A	2/3	154.93 [9.27-2590.28]	0.00045	7	0.51 (0.57)	301	1.10 (0.72)
	rs185763236	Splice donor (-/-)	C/A	1/-	234.86 [1.47-37624.10]	0.035	1	4084.69 (0.0030*)		

**Genotyped, **Genotyped and tested without kinship adjustment. Consequence: Amino acid alteration matching the SIFT score, SIFT=0-1 (deleterious-tolerated), and PolyPhen=0-1 (bening-damaging); if multiple transcript alterations, we report the most severe consequence score, i.e., PolyPhen score may not*

match consequence. MAC=MAC in WGS / MAC in WES, REF=Reference allele, ALT=Alternative allele, OR=Odds ratio, CI=confidence interval.

Figures

Fig. 1: Study design. We studied whole exome- (WES) and whole genome sequencing (WGS) in individuals with T1D. We performed additive and recessive single-variant analyses (MAC≥5) across the genome, using fixed-effects meta-analysis for variants available in both data sets, with a minimal adjustment setting i.e., the calendar year of diabetes onset, sex and two first genomic data principal components, and repeated the analyses with an additional DKD adjustment. We conducted gene aggregate analyses with minimal adjustment separately for protein altering variants (PAVs) and protein truncating variants (PTVs); and repeated the analyses with an additional DKD adjustment. Furthermore, we conducted minimally adjusted intergenic aggregate analyses within genomic windows by statistically up-weighting functionally important and rare variants; and within established enhancers and promoters, by weighting variants according to their rarity. We performed stroke subtype association analyses for lead findings.

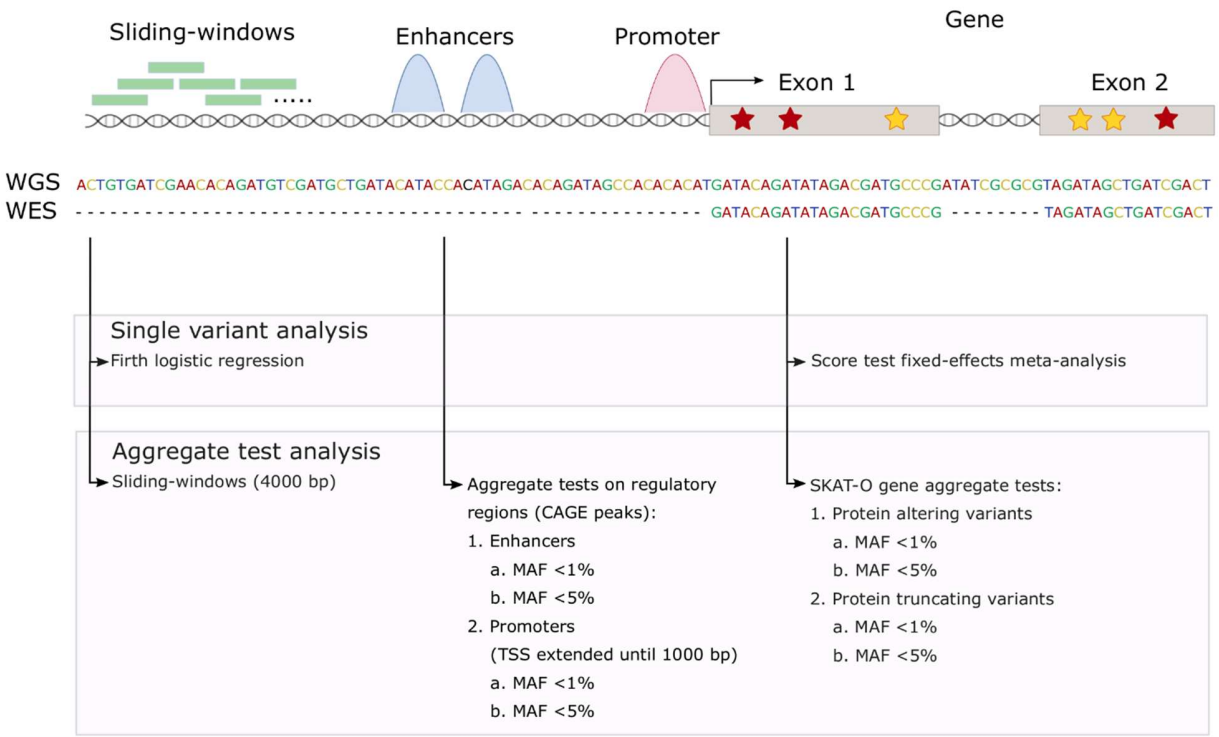


Fig. 2: Additive single variant analysis with minimal adjustment. **A.** Manhattan plot, **B.** LocusZoom plot of 4q33.1-34 (LD structure according to a nearby variant, rs4386563).

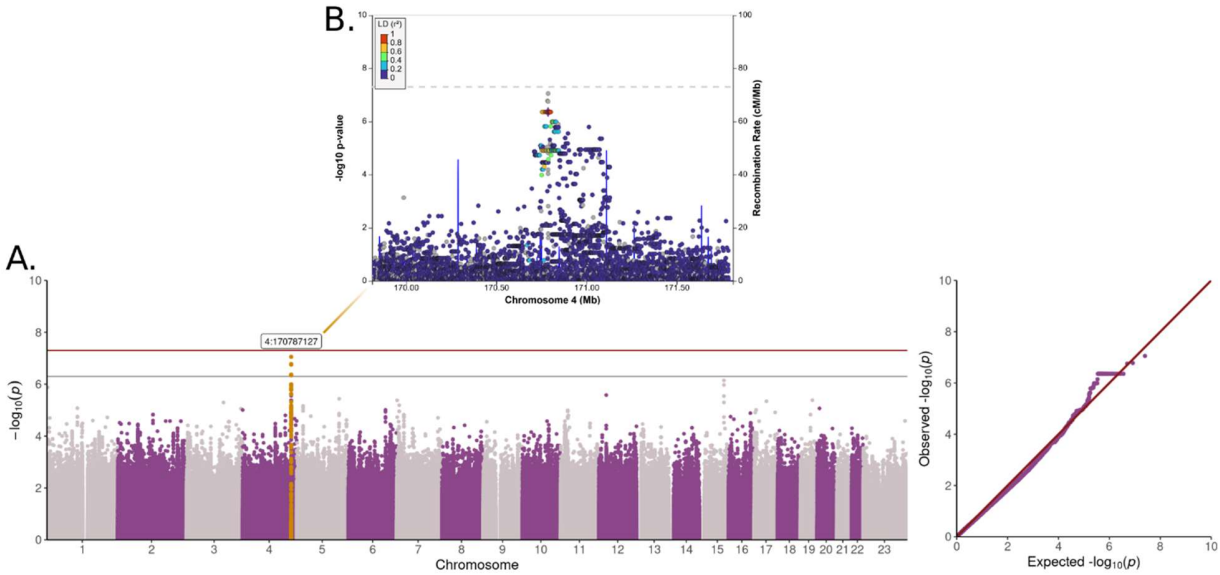


Fig. 3: Discovered genes with the SKAT-O gene aggregate tests. A. Minimal adjustment, **B.** Additional adjustment for diabetic kidney disease. The color indicates the -$\log_{10}$(*p*-value), with darker color indicating more significant finding. Only the rare variant model (MAF≤1%) is reported, if no low-frequency variants (1%<MAF≤5%) were available in the gene. Bonferroni corrected significance thresholds: 4×10^{-6} (protein altering variant, PAV≤1%), 4×10^{-6} (PAV≤5%), 5×10^{-5} (protein truncating variant, PTV≤5%), and 7×10^{-5} (PTV≤1%). Number of variants and cumulative minor allele count (MAC) given based on the combined stroke phenotype.

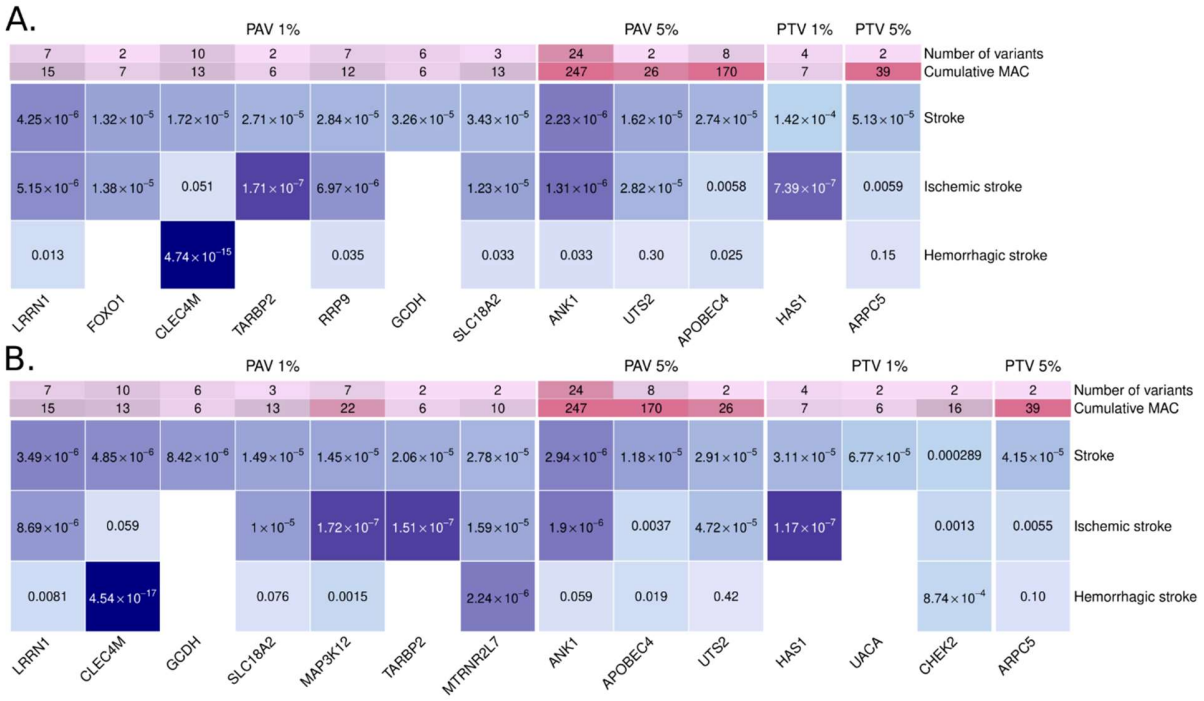


Fig. 4: Variants in significant stroke-associated genes after multiple testing correction (SKAT-O). *Diabetic kidney disease adjusted stroke model.

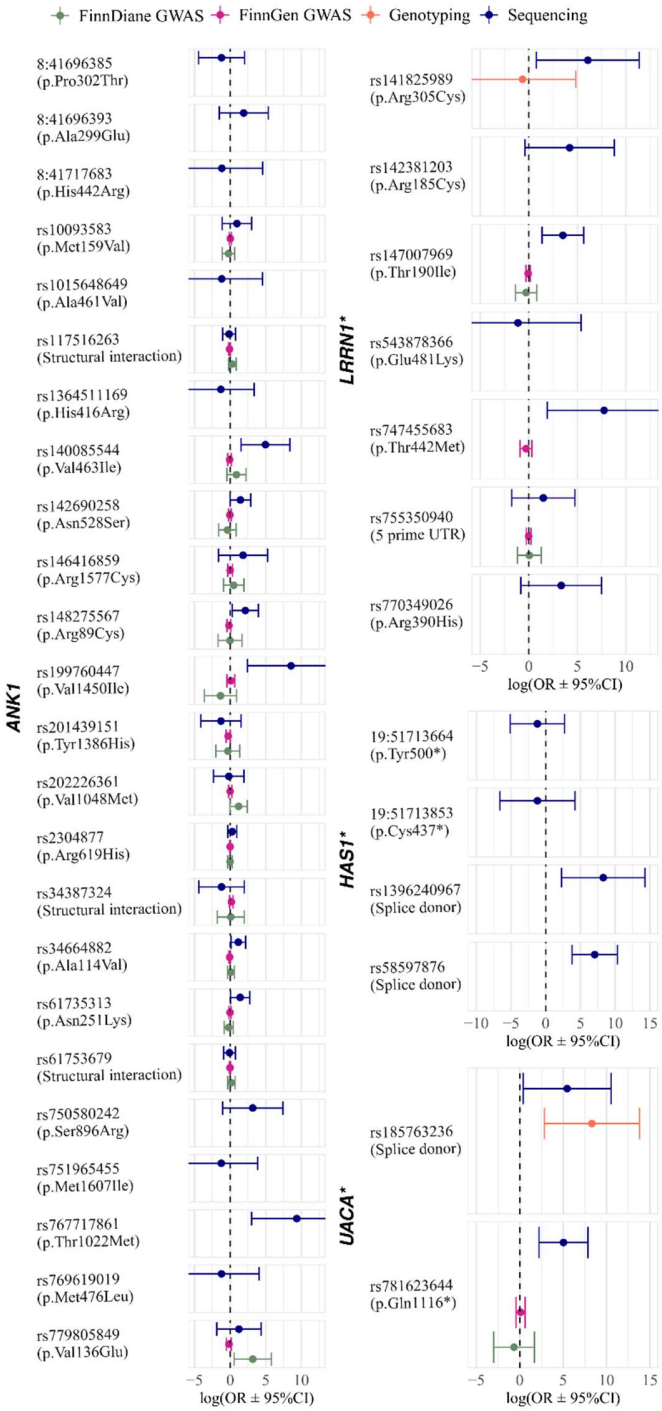
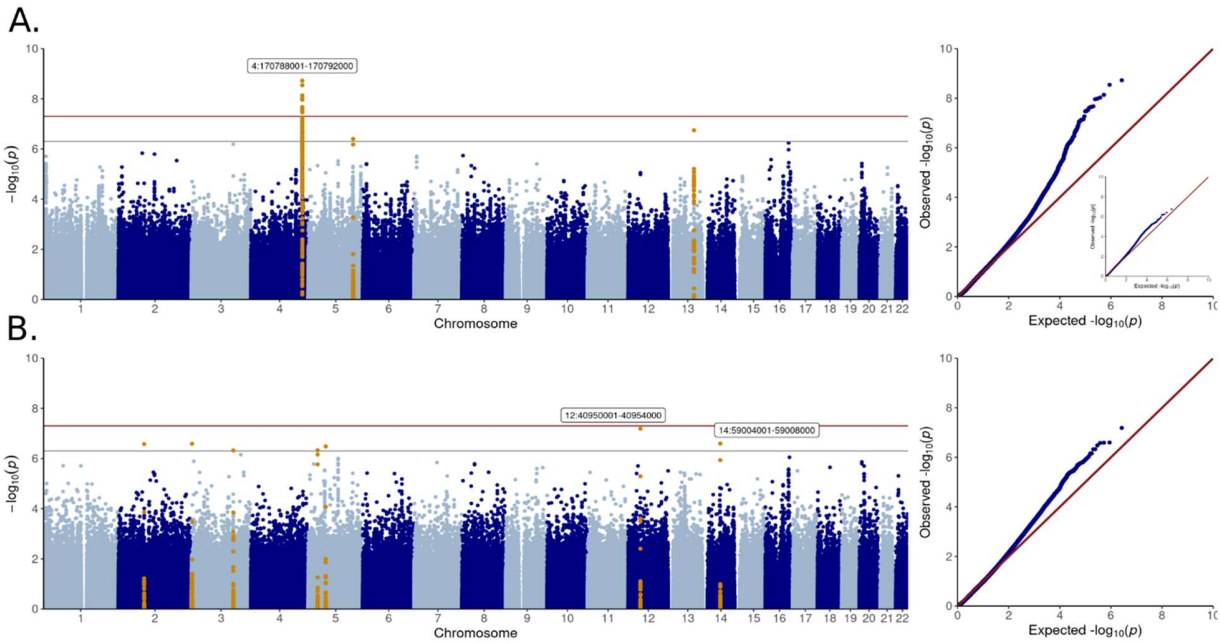


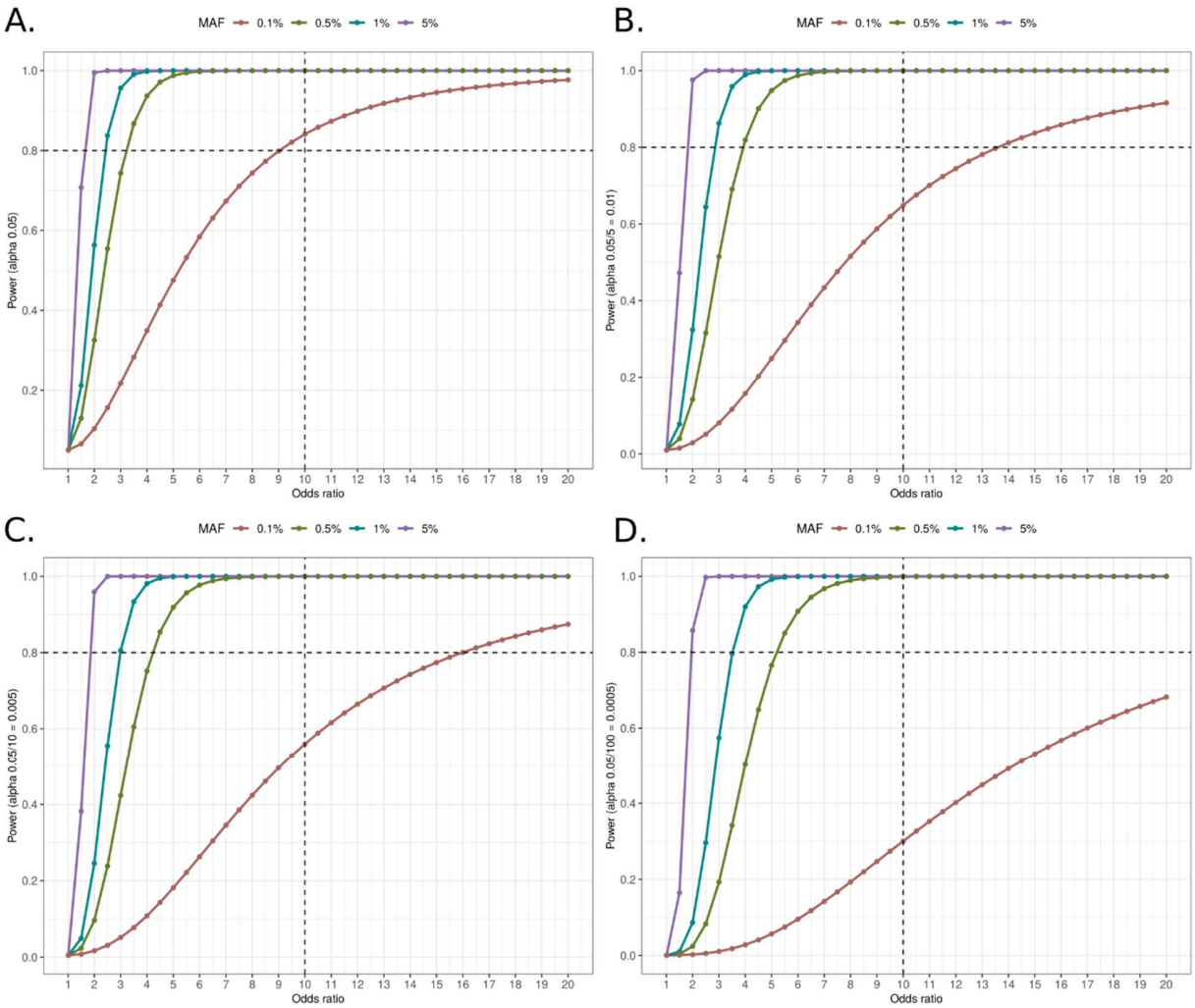
Fig. 5: STAAR sliding-window analyses. A. MAF≤5% (inserted QQ-plot without the 4q33-34.1 region), and **B.** MAF≤1%.



Extended Data Figures

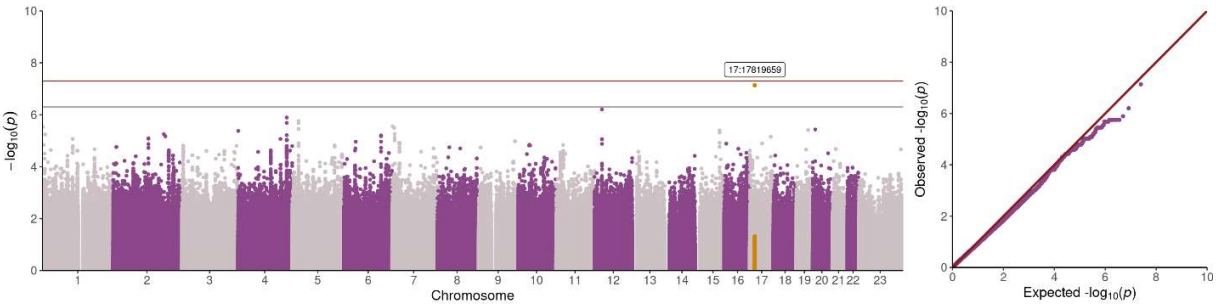
Extended Data Fig 1. Statistical power in replication

Statistical power to replicate low-frequency (MAF=5%, MAF=1%) and rare (MAF=0.5%, MAF=0.1%) variants with the number of patients in FinnDiane GWAS replication (N=3,945, N_{cases}=367) using different alpha levels: **A.** Nominal significance (0.05), **B.** Multiple testing corrected threshold (0.01), **C.** Multiple testing corrected threshold (0.005), **D.** Multiple testing corrected threshold (0.0005).



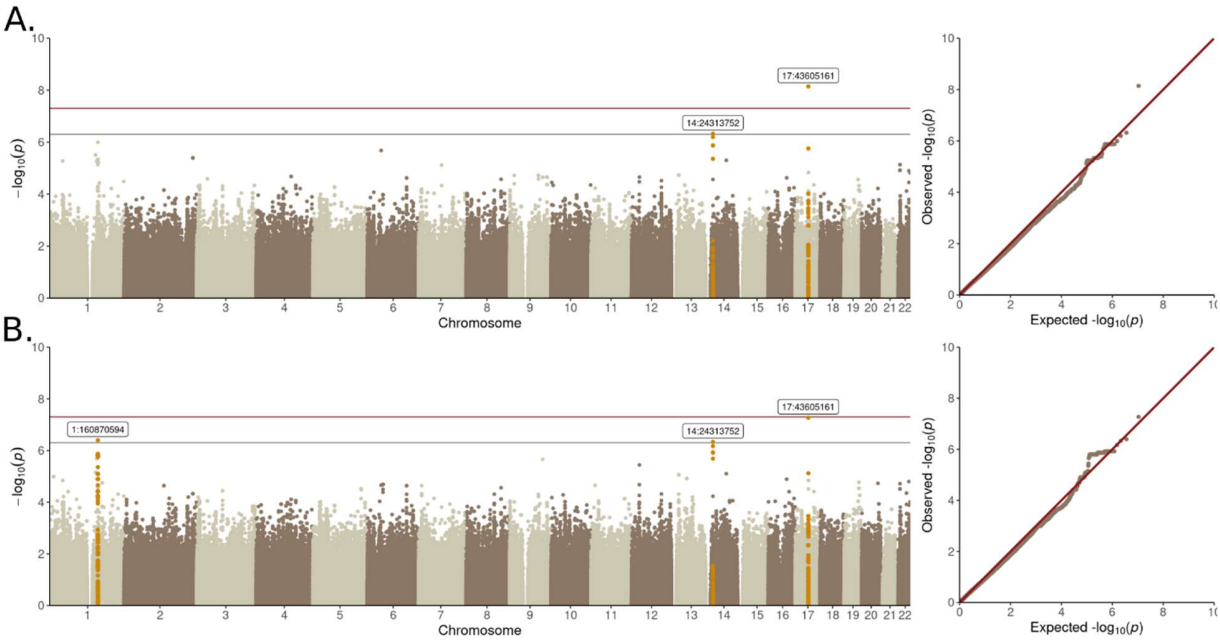
Extended Data Fig 2. Single variant analysis with DKD adjustment

Stroke single variant analysis Manhattan plot (Firth regression or score test fixe-effects meta-analysis).



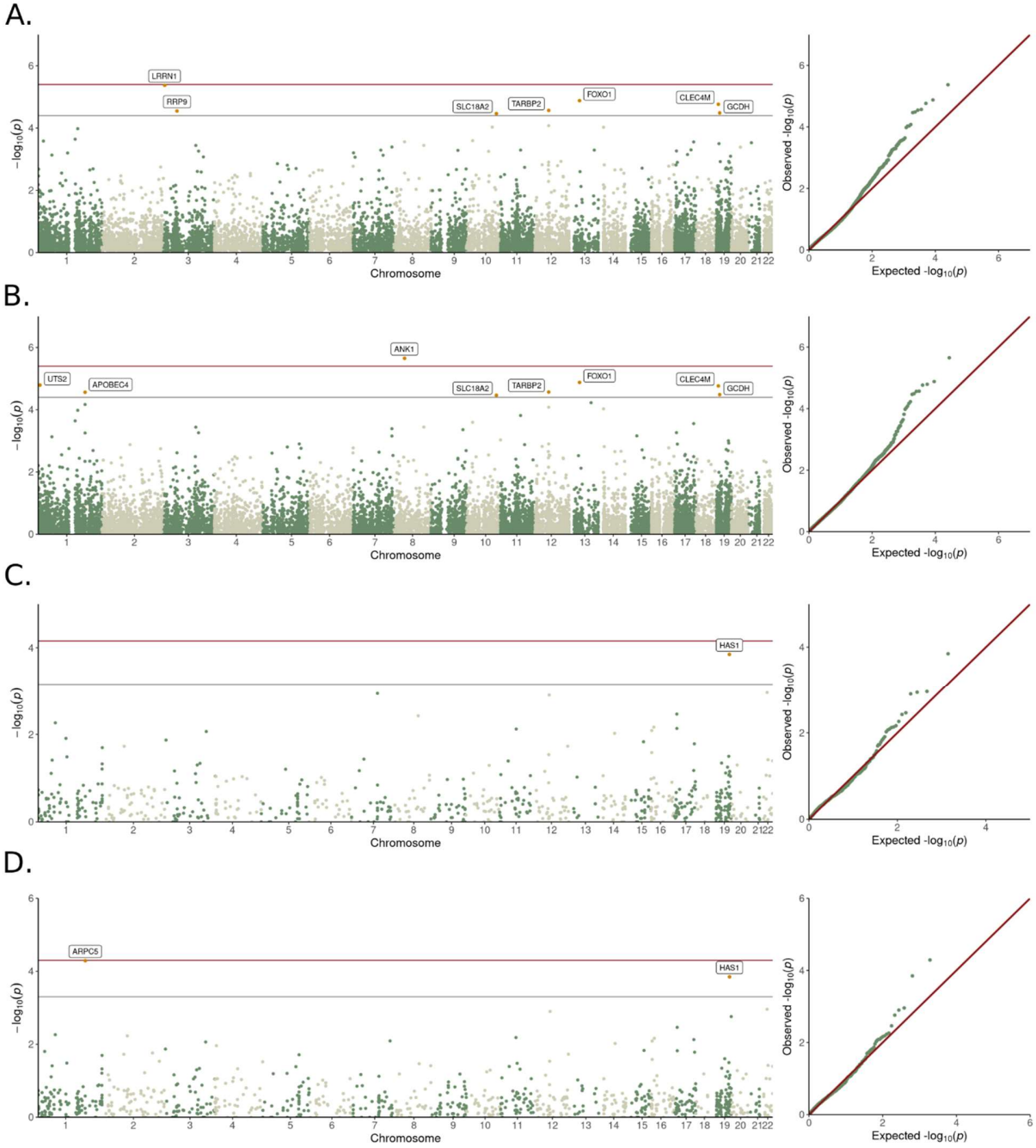
Extended Data Fig 3. Recessive single variant analyses

Stroke recessive autosomal single variant analysis Manhattan plots (Firth regression or score test fixed-effects meta-analysis): **A.** Minimal adjustment, and **B.** An additional DKD adjustment.



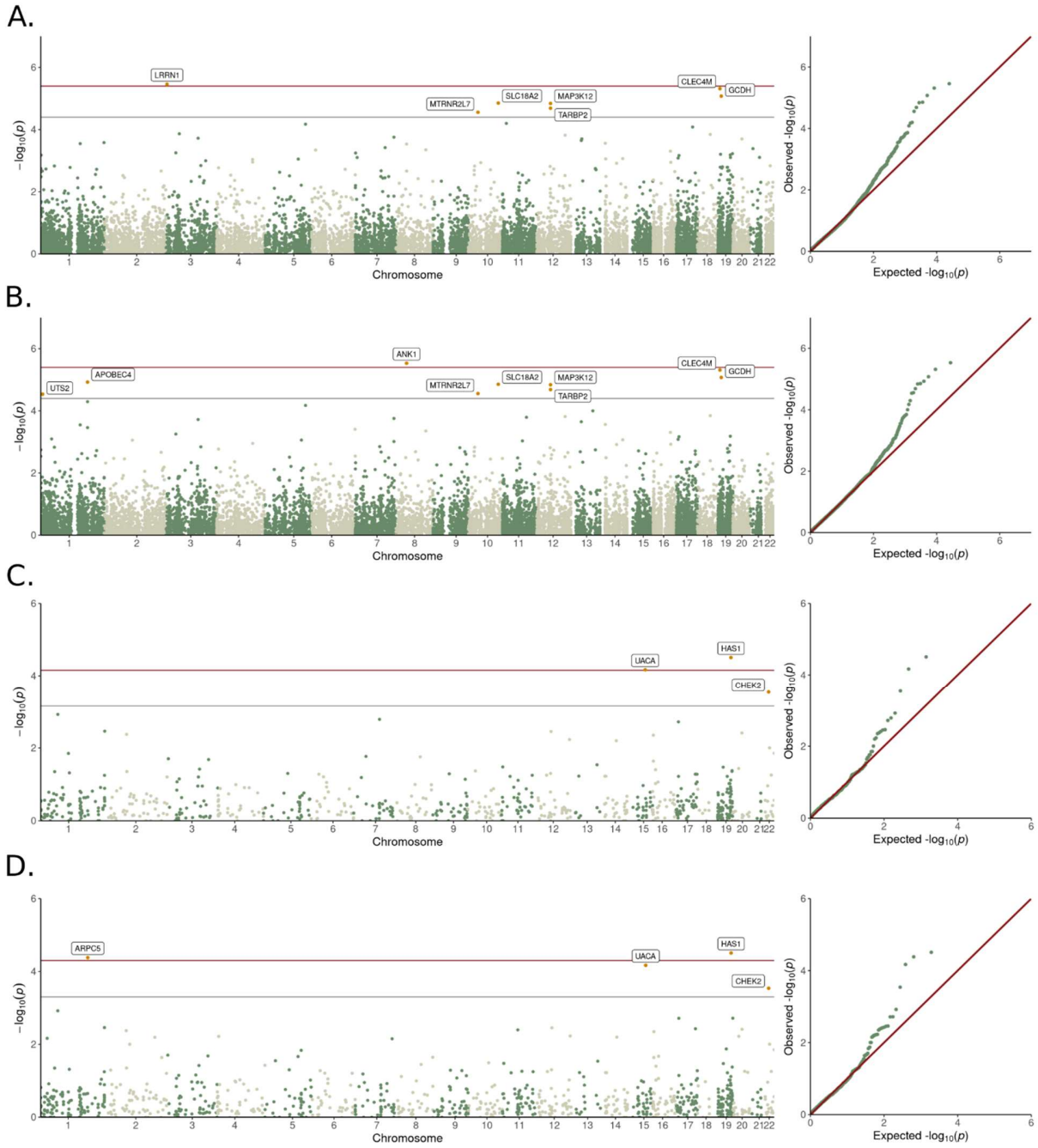
Extended Data Fig 4. SKAT-O minimal model Manhattan plot

A. Protein altering variant (PAV) ≤1%, **B.** PAV≤5%, **C.** Protein truncating variant (PTV) ≤1%, **D.** PTV≤5%.



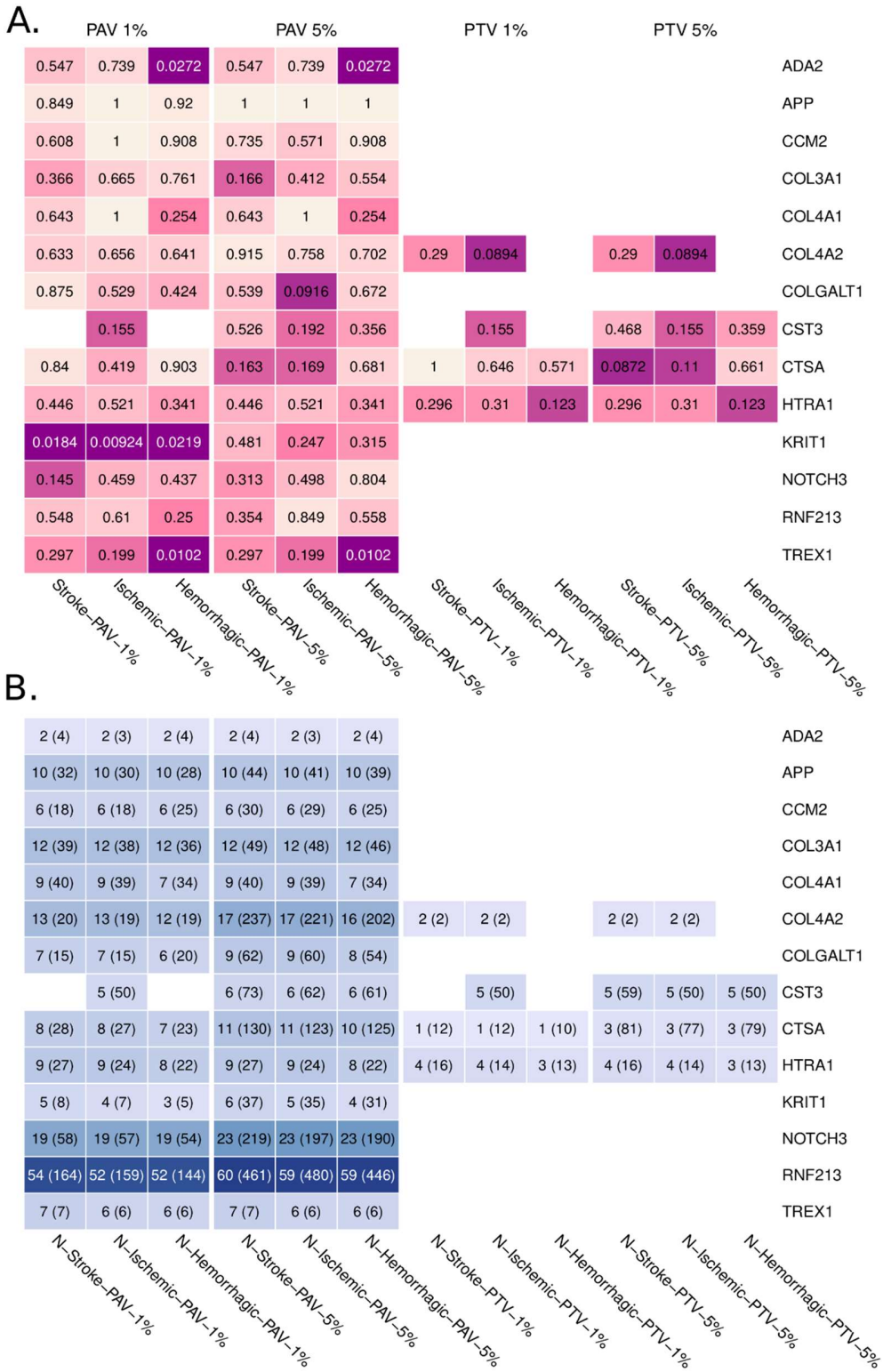
Extended Data Fig 5. SKAT-O with additional DKD adjustment Manhattan plot

A. Protein altering variant (PAV) ≤1%, **B.** PAV≤5%, **C.** Protein truncating variant (PTV) ≤1%, **D.** PTV≤5%.



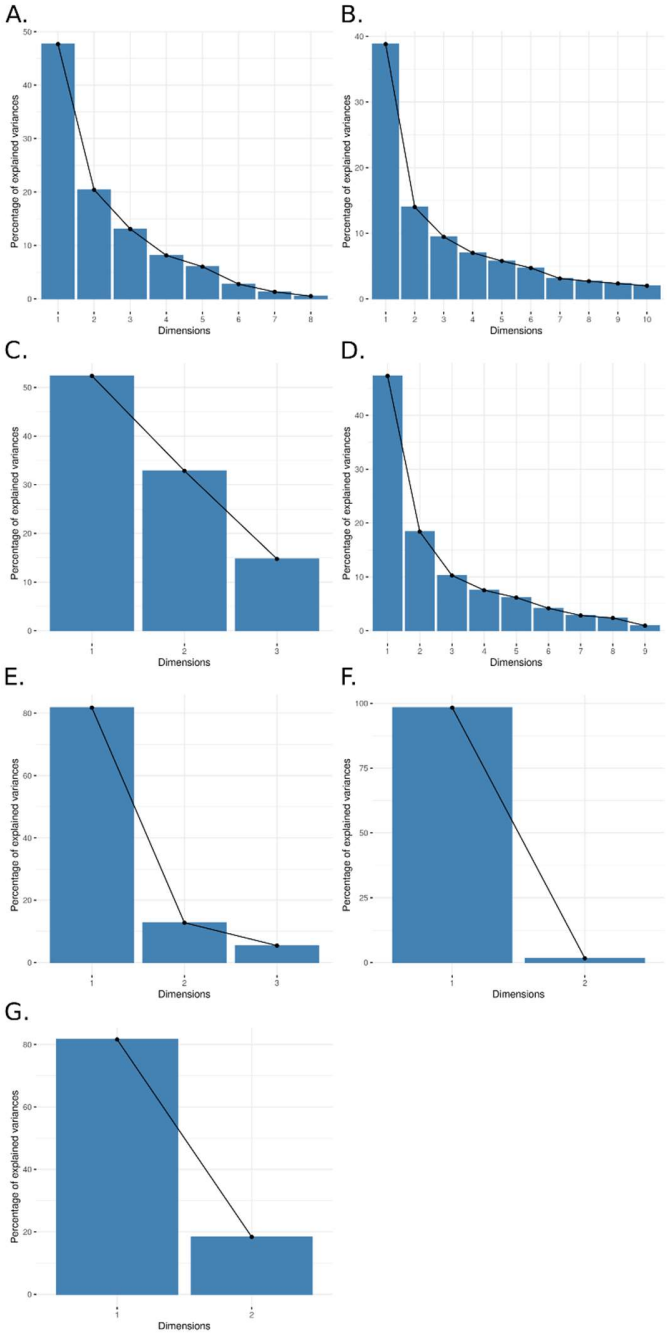
Extended Data Fig 6. Known Mendelian stroke-risk genes in T1D

A. SKAT-O *p*-value, **B.** $N_{variant}$ (MAC) in SKAT-O. Results presented with minimal adjustment.



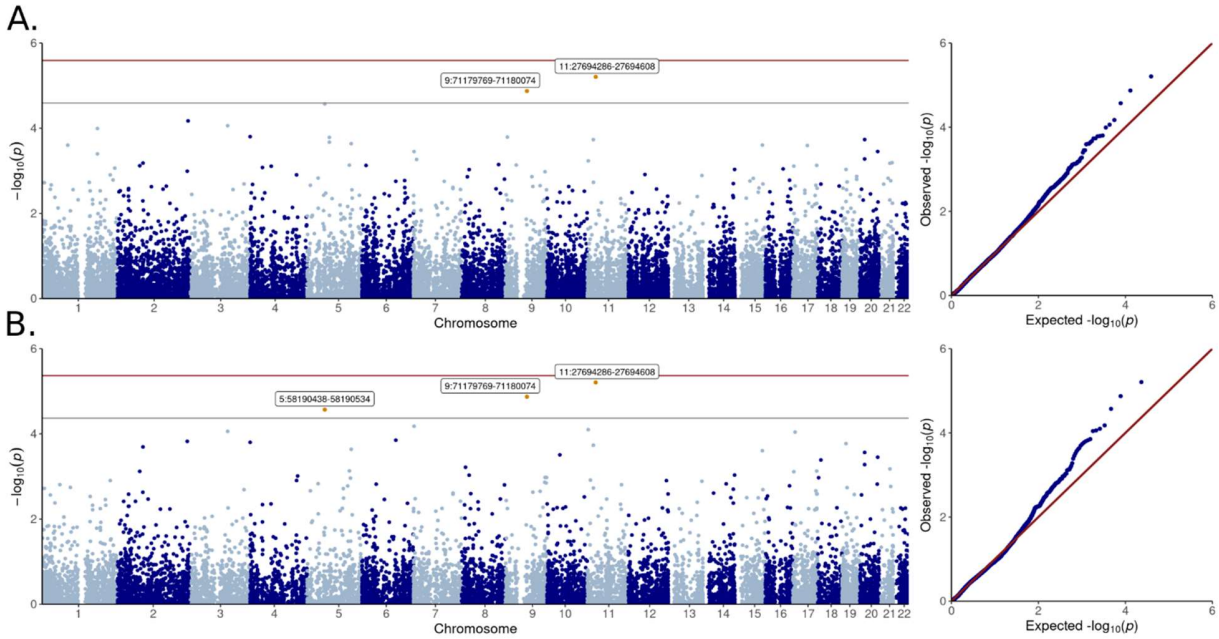
Extended Data Fig 7. Annotation PCA within functional classes

A. Conservation, **B.** Epigenetics, **C.** microRNA, **D.** Mutation density, **E.** Protein function, **F.** Transcription factor, **G.** Proximity to transcription start- and end sites. Variance explained by the corresponding annotation principal component (aPC) is presented at Dimension 1.



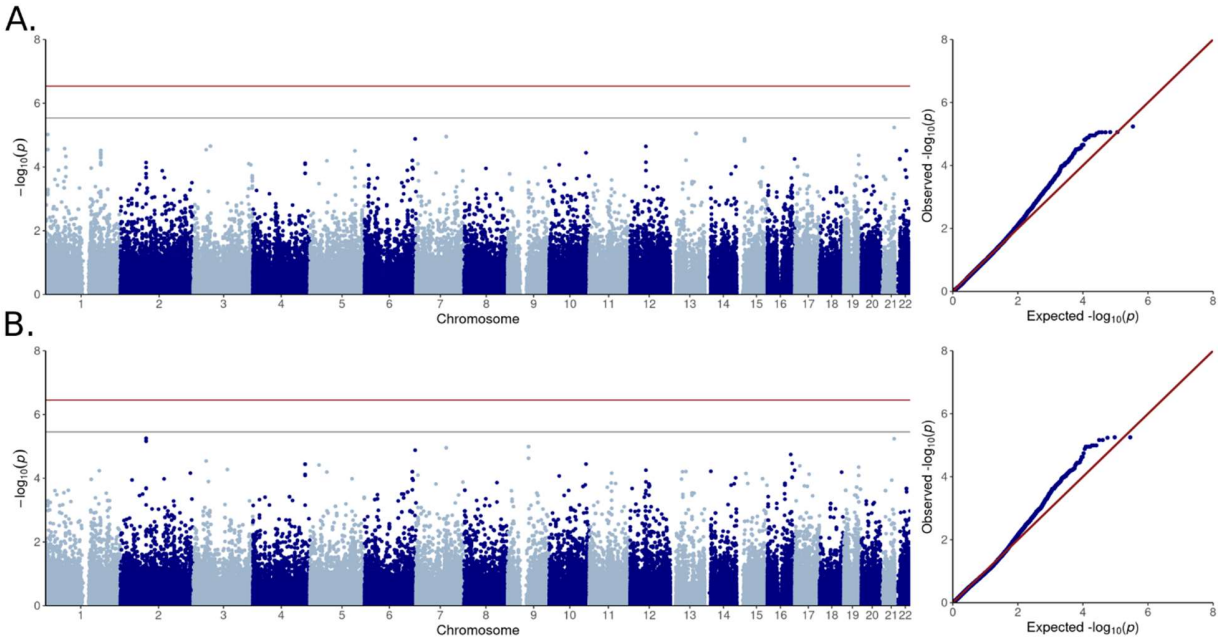
Extended Data Fig 8. Stroke enhancer association Manhattan plot

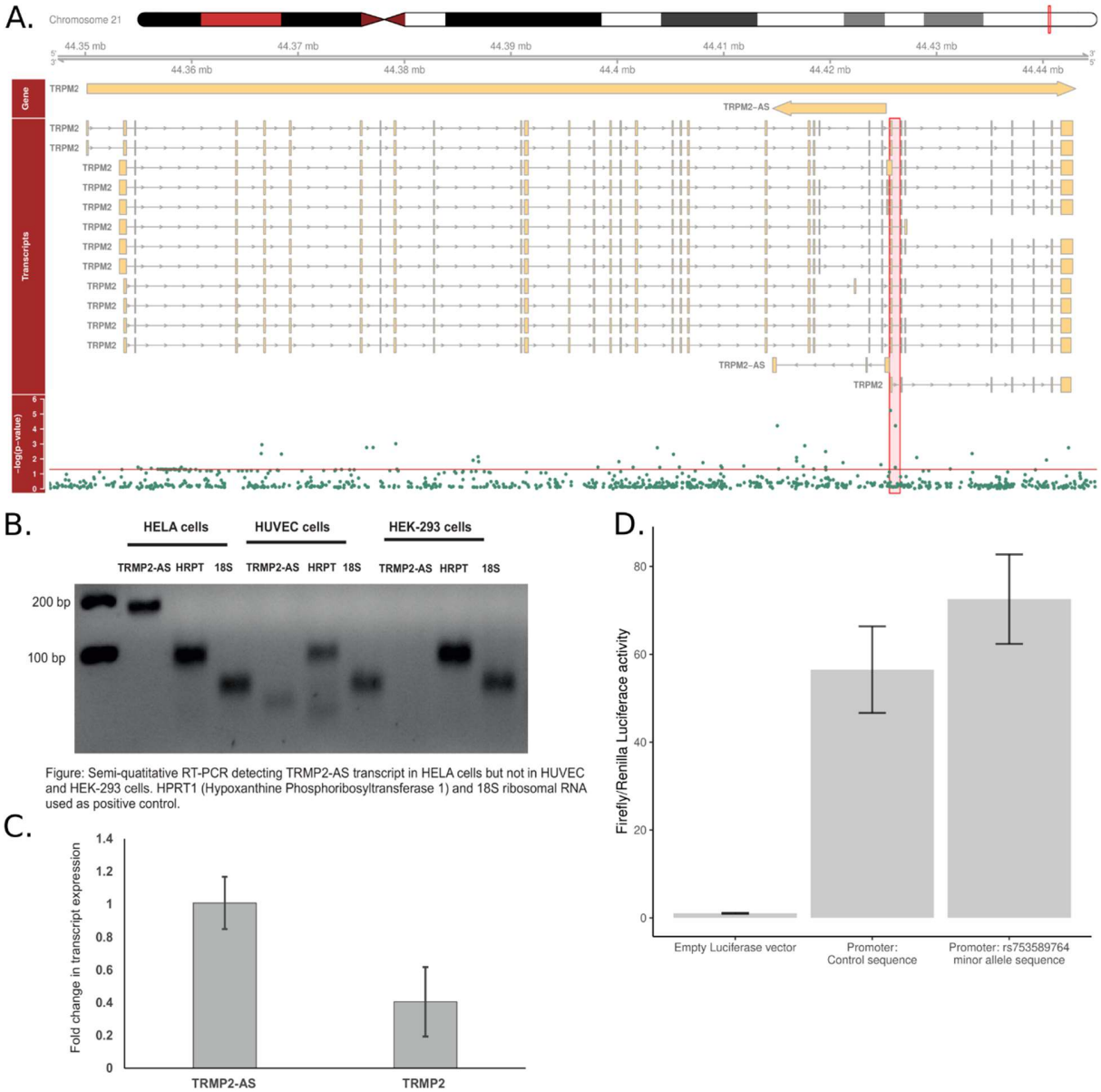
Variants weighted with PHRED scale minor allele frequencies. **A.** MAF≤5%, and **B.** MAF≤1%.



Extended Data Fig 9. Stroke promoter association Manhattan plot

Variants are weighted with PHRED scale minor allele frequencies. **A.** MAF≤5%, and **B.** MAF≤1%.



Extended Data Fig 10. *TRPM2-AS* regional plot and experimental data

A. Regional plot of *TRPM2* and *TRPM2-AS* extended region; the discovered promoter is highlighted (red). **B.** *TRPM2-AS* expression in HELA, HEK-293 and HUVEC cell lines, **C.** Relative expression of *TRPM2-AS* and *TRPM2* transcripts in HELA cells (hypoxanthine phosphoribosyltransferase 1 used as reference transcript to normalize quantitative RT-PCR), **D.** Firefly/Renilla luciferase assay of promoter activity. Empty vector (mean=1.0, 12 technical repeats) as transfection control for baseline luciferase activity was compared to *TRPM2-AS* control promoter, i.e., major allele in all identified *TRPM2-AS* variants (mean=56.5, 11 technical repeats, *p*-value=0.00022); which was further compared to a *TRPM2-AS* promoter with rs753589764 minor allele (mean=72.6, 11 technical repeats, *p*-value=0.27): All cloned before firefly reporter gene to evaluate potential transcriptional promoter activity. Statistical significance was assessed with Student's t-test and error bars represent standard error.