1 Title:

2 De novo mutation hotspots in homologous protein domains identify function-altering

- 3 mutations in neurodevelopmental disorders
- 4
- 5 Authors and affiliations:
- 6 Laurens Wiel^{1,2,3}*, Juliet E. Hampstead¹*, Hanka Venselaar², Lisenka E. L. M. Vissers⁴, Han
- 7 G. Brunner¹, Rolph Pfundt⁴, Gerrit Vriend⁵, Joris A. Veltman⁶, and Christian Gilissen¹
- 8
- 9 1. Department of Human Genetics, Radboud Institute for Molecular Life Sciences, Radboud
- 10 University Medical Center, Nijmegen, 6525 GA, the Netherlands
- 11 2. Centre for Molecular and Biomolecular Informatics, Radboud Institute for Molecular Life
- 12 Sciences, Radboud University Medical Center, Nijmegen, 6525 GA, the Netherlands
- 13 3. Present address: Department of Medicine, Division of Cardiovascular Medicine, School of
- 14 Medicine, Stanford University, Stanford, CA, USA
- 15 4. Department of Human Genetics, Donders Institute for Brain, Cognition and Behaviour,
- 16 Radboud University Medical Center, Nijmegen, 6525 GA, the Netherlands
- 17 5. Baco Institute of Protein Science, Baco, 5201, Mindoro, Philippines.
- 18 6. Biosciences Institute, Faculty of Medical Sciences, Newcastle University, Newcastle upon
- 19 Tyne, NE1 3BZ, United Kingdom
- 20
- 21 *: These authors contributed equally.
- 22 Corresponding author(s): Christian Gilissen

23 Abstract

24 Variant interpretation remains a major challenge in medical genetics. We developed Meta-25 Domain HotSpot (MDHS) to identify mutational hotspots across homologous protein 26 domains. We applied MDHS to a dataset of 45,221 de novo mutations (DNMs) from 31,058 27 patients with neurodevelopmental disorders (NDDs) and identified three significantly 28 enriched missense DNM hotspots in the ion transport protein domain family (PF00520). The 29 37 unique missense DNMs that drive enrichment affect 25 genes, 19 of which were 30 previously associated with NDDs. 3D protein structure modelling supports function-altering 31 effects of these mutations. Hotspot genes have a unique expression pattern in tissue, and we 32 used this pattern alongside in silico predictors and population constraint information to 33 identify candidate NDD-associated genes. We also propose a lenient version of our method, 34 which identifies 32 hotspot positions across 16 different protein domains. These positions are 35 enriched for likely pathogenic variation in clinical databases and DNMs in other genetic 36 disorders.

37

38 Key Words

- 39 Mutational hotspot detection; homologous protein domains; disease gene identification; de
- 40 novo mutations; developmental disorders; variant interpretation; pathogenicity

41 Introduction

42 The interpretation of sequence variation in the context of disease remains one of the biggest 43 challenges in genetics. De novo mutations (DNMs) in protein-coding genes are an established cause of neurodevelopmental disorders (NDDs)¹, and roughly $\sim 45\%$ of NDDs are caused by 44 a DNM in a protein-coding gene.^{2,3} By modelling the probability of DNMs occurring in 45 specific genes, one can identify genes that are enriched for DNMs in patient cohorts, 46 47 provided that large enough cohorts are available. This statistical identification of NDDassociated genes requires ever-larger patient collections.³⁻⁷ A recent study of DNMs in 48 49 31,058 patients with NDDs concluded that NDD-association of genes still is far from 50 saturated and that over a thousand NDD-associated genes are still to be identified.⁷

51

52 Several studies of NDD patients have found that for specific genes, missense DNMs cluster in functional regions, and that this fact can be used to identify disease-associated genes.⁷⁻⁹ 53 Conserved protein domains are of particular interest, because they harbour $\sim 71\%^{10}$ of all 54 curated disease-causing missense variants in the Human Gene Mutation Database (HGMD)¹¹ 55 and ClinVar¹². Indeed, missense DNMs in NDD genes are almost three times more likely 56 located in protein domains.⁷ Clustered missense DNMs in these genes may not act through 57 58 haploinsufficiency, but rather through dominant negative or gain-of-function effects.^{8,9} The detection of mutation clusters, or hotspots, can be a crucial step towards associating genes 59 60 with NDDs¹³ and for gaining insight into underlying disease mechanisms.⁹

61

62 Aggregation of variation across homologous domains can be a useful method to gain insight into patterns of variation^{10,14,15} and therefore can also increase statistical power to detect 63 hotspot positions. Methods such as mCluster¹⁶ and the DS-Score¹⁷ have been developed to 64 detect re-occurrence of missense mutations at equivalent positions in protein domains to 65 66 identify passenger mutations in cancers. In line with the same principle, we developed Meta-67 Domain HotSpot (MDHS), a novel method to detect mutation clustering at evolutionary 68 equivalent positions across homologous protein domains and applied this method to DNMs 69 from a large cohort of NDD patients to identify protein consensus positions enriched for 70 missense variation.

72 **Results**

73 General description of the data and the processing steps

To identify hotspots of de novo mutations in homologous protein domains, we computed 74 75 MDHS (Equation 1) based on unique DNMs in NDD patients (Figure 1). This was done for 76 each variant type separately (missense, synonymous, nonsense). We first mapped 45,221 DNMs resulting from 31,058 patients with developmental disorders⁷ onto gene transcripts 77 78 using MetaDome¹⁴ (Methods). Then, we aggregated these to 12,389 meta-domain¹⁰ positions (Supplementary Data S1). The final 15,322 DNMs represent 73.7% missense (n=11,288), 79 80 21.1% synonymous (n=3,229), and 5.3% stop-gained mutations (n=805) (Supplementary 81 Table 1). 82

83 Stringent DNM hotspots identified using MDHS

84 We initially used a stringent approach where recurrent DNMs were counted only once to 85 prevent hotspots driven by DNMs in a single gene. Using all 11,288 missense DNMs in 2,032 86 protein domain families, our method identified three significant hotspots (Supplementary 87 Data S2-S4) comprising 37 unique missense DNMs (57 total) in 25 different genes 88 (Supplementary Data S2). Strikingly, all three hotspots are in domains belonging to the Ion 89 Transport protein domain family (PF00520) (Figure 2). To validate our approach, we also 90 performed the stringent method separately for the 3,229 synonymous and 805 stop-gained 91 DNMs in our cohort and identified no significant hotspots (Supplementary Data S5-S6).

92

93 The three significant missense hotspots we identify are located on domain consensus positions p.96 (10 unique DNMs, $p = 3.6 \times 10^{-2}$), p.102 (13 unique DNMs, $p = 7.1 \times 10^{-5}$), 94 and p.231 (14 unique DNMs, $p = 7.5 \times 10^{-6}$) of the ion transport domain family. The Ion 95 96 Transport protein domain family is one of four protein domain families that we previously 97 found to be significantly enriched with missense DNMs in NDD-associated genes.⁷ Of the 25 98 'hotspot genes' identified with a missense DNM at a hotspot, 19 are known NDD-associated 99 genes, representing a 3.17-fold enrichment of known NDD-associated genes ($p = 1.11 \times 10^{-13}$ 100 chi-square test, **Supplementary Table 3**). The remaining 6 genes have not yet been 101 associated with NDDs (Table 1).

102

103 Effects of missense mutations at stringent hotspots on protein structure

104 Mutations that cluster in genes have previously been associated with likely function-altering 105 effects.⁹ We find that 6 out of 16 hotspot genes present in the Developmental Disorder 106 Genotype2Phenotype (DDG2P) gene list are known to have an activating or gain-of-function 107 mutation consequence (p = 0.0008, Fisher's exact test), underscoring that missense mutations 108 at hotspot positions are likely function-altering (Supplementary Table 4). To further 109 investigate this, we created 3D protein structure homology models for each of the 25 genes 110 (Supplementary Data S1, Methods). Ion transport protein domain 3D structures are 3-fold 111 more identical to each other in conformation than their protein sequences would suggest (CATH-Gene3D ID: 1.20.120.350).¹⁸ This structural overlap encouraged us to investigate 112 113 whether molecular effects of missense variants at these hotspots are likely to have similar 114 impact on domain function across the 25 genes (Supplementary Data S8).

115

In the 25 homology models we found that hotspot p.96 (**Figure 3A**) and p.102 (**Figure 3B**) are part of the voltage-sensing helix that is important for the channel (in-)activation.¹⁹ These results are in line with functional studies that have been performed for missense mutations at two of these hotspots.^{20,21} Hotspot p.231 (**Figure 3C**) is part of the channel gate at the end of a transmembrane helix (**Supplementary Data S8**). In addition, we found that missense mutations follow a specific pattern for each of these hotspots. Of the 13/16 missense DNMs

122 located at hotspot p.96 and 20/20 at p.102 change the positively charged wild-type residue to 123 lose the positive charge. Losing positive charges at these locations has previously been described to trigger a function altering disease-mechanism (Figure 3AB).^{20,21} At hotspot 124 p.231 20/21 of the missense DNMs changes the wild-type residue from a small into a larger 125 126 residue. This change in residue size likely impacts the pore closure. This hypothesis is shared 127 by Kortüm et al. who suggest this likely causes a steric hindrance and result into a functionaltering mechanism of disease (Figure 3C).²² Overall, this shows that missense mutations at 128 129 the identified hotspots are likely deleterious to domain function.

130

131 Stringent hotspot genes are constrained against missense and loss of function variation

132 Dominant NDD genes are characterized by population constraint against damaging genetic variation.^{23,24} We compared observed counts of loss of function, missense, and synonymous 133 variants in control, NDD-associated, hotspot and proposed novel hotspot genes in gnomAD 134 v2 to expected counts based on a null mutational model.²⁵ Both hotspot and novel hotspot 135 136 genes are constrained against loss of function and missense variation (Figure 4A). Novel 137 hotspot genes have lower constraint against loss of function variation than hotspot genes 138 (Supplementary Data S9). We also considered whether hotspots were found in regions of 139 particular constraint against missense variation within genes. In total, 2,700 genes have 140 statistical evidence of regional differences in missense constraint.²⁵ Of these, 16 are hotspot 141 genes, representing a significant enrichment compared to control genes (Fisher's exact test p 142 $< 2.2 \text{ x } 10^{-16}$) and NDD-associated genes (Fisher's exact test p = 0.002, Supplementary 143 Figure 1). Three are proposed novel hotspot genes (KCNH5, CACNA1B, TPCN1). Using 144 regional missense constraint information, we show that PF00520 domains in hotspot genes 145 are significantly more constrained against missense variation than PF00520 domains in control genes ($p = 1.4 \times 10^{-7}$, Wilcoxon rank-sum test; Figure 4B), but similarly constrained 146 147 compared to NDD-associated genes without a hotspot that also contain a PF00520 domain (p 148 = 0.65, Wilcoxon rank-sum test).

149

150 Brain-specific expression of stringent hotspot genes

We analysed the expression of the 19 known hotspot genes in approximately 948 donors 151 152 across 54 tissues from the GTEx v8 release.²⁶ We observed that NDD genes and control genes have distinct gene expression patterns, with a higher proportion of NDD genes 153 constitutively expressed across all tissues (p < 2.2×10^{-16} , Fisher's exact test; 154 155 **Supplementary Table 5**). Hotspot genes share a characteristic expression pattern compared 156 to these two groups (Figure 5A), with a significantly higher proportion of hotspot genes 157 expressed in the brain compared to control genes and significantly lower proportion 158 expressed in all other tissues compared to NDD genes (in 40/42 non-brain tissues, 159 **Supplementary Data S10**). Given this tissue-specific expression pattern, we grouped GTEx 160 tissues into two tissue groups (brain and other tissues, **Methods**). The hotspot gene set is 161 significantly enriched for genes with higher expression in brain compared to control genes (89.4% versus 19.8% expressed higher in brain, $p = 2.985 \times 10^{-5}$, Fisher's exact test) and 162 163 NDD genes (89.4% versus 31.3%, p = 0.002, Fisher's exact test) (Supplementary Figure 2). 164 Only two hotspot genes do not have higher expression in brain: SCN10A, which is 165 constitutively unexpressed across tissues in GTEx samples, and CACNA1C (median TPM in 166 brain = 2.94, median TPM in other tissues = 4.35). We further show that this expression 167 pattern is not characteristic of all genes containing an ion transport domain, but only the 168 subset of these genes statistically associated to NDDs (Supplementary Data S11, 169 Supplementary Figure 3-4).

171 We also compared the TPM distribution in brain and other tissues for control genes, NDD-172 associated genes, hotspot genes, and the six novel hotspot genes (Methods; Figure 5B). Both 173 hotspot and NDD-associated genes had significantly higher TPM in brain tissues than control genes (p = 0.0039 and p < 2.2×10^{-16} , Wilcoxon rank-sum test), and both hotspot genes and 174 175 control genes had significantly lower TPM in other tissues than NDD-associated genes (p < p2.2 x 10^{-16} and p = 2.08 x 10^{-8} , Wilcoxon-rank sum test; Supplementary Figure 5). 176 177 Modelling suggests that CACNA1B and KCNG1 belong to the hotspot gene distribution by 178 odds ratio (Supplementary Data S12). Additionally, we find 6 NDD-associated PF00520 179 domain-containing genes (HCN1, TRPV3, KCNO5, KCNC1, KCNC3, TRPM3) that also 180 belong to the hotspot gene distribution by odds ratio. We hypothesize that missense mutations 181 at hotspot positions in these 6 genes may also cause NDDs.

182

183 Lenient DNM hotspots identified using MDHS

184 We also implemented a lenient version of MDHS that counts all missense variants at protein 185 consensus positions, even if they recur between individuals (see Methods, Supplementary 186 Figure 6). Counting recurrent missense variants gives us more power to detect hotspot 187 positions, but these hotspot positions may be driven by missense variation in a single domain. 188 We detected 32 lenient missense DNM hotspots across 16 Pfam protein domain families 189 using this method (Supplementary Table 2). We find a significant enrichment of genes statistically associated to NDD (Fisher's exact $p < 2.2 \times 10^{-16}$) and DDG2P genes (Fisher's 190 exact $p < 2.2 \ge 10^{-16}$) at lenient hotspot positions (Supplementary Figure 7). 191

192

193 We also find that missense variants at these positions are significantly more likely to be 194 pathogenic or likely pathogenic in clinical databases (VKGL, Figure 6A; ClinVar, Figure 195 **6B**). We compared the proportion of reported likely pathogenic (LP) missense variants at 196 hotspot positions to those at other protein consensus positions across the 16 Pfam domain 197 families with a lenient hotspot (Figure 6AB). We find a significant enrichment of LP variants at hotspot positions when we consider all positions (Fisher's exact $p < 2.2 \times 10^{-16}$, ClinVar; p 198 $< 2.2 \text{ x} 10^{-16}$, VKGL), only positions without a DNM in our cohort (Fisher's exact p < 2.2 x199 10^{-16} , ClinVar; $p < 2.2 \times 10^{-16}$, VKGL), and only codons without a DNM in our cohort 200 (Fisher's exact $p < 2.2 \ge 10^{-16}$, ClinVar; $p = 3.08 \ge 10^{-13}$, VKGL; Supplementary Table 6). 201

202

203 Lenient hotspots are enriched for missense variation in autism-spectrum disorders

We investigated if we could find further evidence for the identified lenient hotspots in a combined cohort of publicly available *de novo* mutation datasets for autism-spectrum disorders (ASD; 11,986 ASD probands, 35,584 individuals) and congenital heart defects (CHD; 2,654 trios) alongside DNMs from unaffected individuals (1,740 ASD siblings, 1,548 population controls; see **Methods**).

209

210 We observe a significant enrichment of missense DNMs at hotspot positions in NDD and 211 ASD cohorts compared to unaffected individuals (Fisher's exact $p = 3.5 \times 10^{-13}$, NDD; 212 Fisher's exact p = 0.007, ASD; Fisher's exact p = 0.07, CHD; Figure 7A, Supplementary 213 **Table 7**). We observe no significant enrichment in synonymous DNMs at hotspot positions in 214 any cohort (Supplementary Table 8). However, we predict that some lenient hotspot 215 positions are driven by mutational processes or ascertainment bias in particular genes and not 216 necessarily by the cumulative effect of pathogenic mutations across several genes. To correct 217 for this, we also tested for an enrichment of missense variants unique to ASD and CHD 218 probands at lenient hotspots (Figure 7B, Supplementary Table 9). We found a significant 219 enrichment of these unique missense variants in ASD probands (Fisher's exact p = 0.047) but 220 not in CHD probands (Fisher's exact p = 1). The majority (10/13, 77%) of the missense

- 221 variants driving this enrichment in ASD probands are in genes statistically associated to
- 222 NDDs.⁷

224 **Discussion**

By exploiting homology within the human genome we were able to identifyied mutational clustering of DNMs at evolutionarily conserved positions across genes that share protein domains. We identify identified three stringent (p.96, p.102, p.231) and 32 lenient mutational hotspots across 16 Pfam domain families using our MDHS method. Missense DNMs at stringent hotspots are located in 25 genes within our cohort. Structural and functional work by us and others suggest that missense mutations at these positions may be gain-offunction.^{27,28}

232

233 The hotspots we statistically identify in our cohort may have broader clinical relevance. We 234 hypothesize that the 19 hotspot genes are examples of a broader class of ion transport 235 domain-containing genes, and that missense mutations at hotspot positions in these genes are 236 generally damaging. Our finding that clinical databases contain many pathogenic missense 237 mutations at hotspot positions in other monogenic disease genes (Supplementary Data S13) 238 support this idea. Other studies have shown that missense mutations in ion transport domaincontaining genes may have position-specific functional effects.^{29,30} Several of these mutations 239 240 occur in genes not statistically associated to NDDs, indicating that missense mutations at 241 hotspot positions could be pathogenic across a variety of disorders. In line with this, we 242 observe that some PF00520 domain-containing NDD-associated genes have lower expression 243 in brain but have a similar level of tissue-specific expression in a non-brain tissue. SCN4A, 244 for example, is not expressed in brain and is predominantly expressed in skeletal muscle. 245 Although the expression pattern of SCN4A is different from the hotspot genes presented in 246 our analysis, hotspot positions in SCN4A are similarly constrained against missense variation 247 (Figure 5B). We hypothesise that phenotypes associated with pathogenic mutations at 248 hotspot positions may vary depending on where the mutated gene is expressed. For example, 249 of the four PF00520 domain-containing genes predominantly expressed in skeletal muscle 250 (SCN4A, CACNA1S, RYR1, and KCNA7), three of these (SCN4A, RYR1, and CACNA1S) have 251 pathogenic missense variation at hotspot positions in clinical databases (Supplementary 252 **Data S13**). Patients with these mutations present with disorders of the skeletal muscle, 253 including myotonia, paramyotonia and hyperkalemic paralysis. Our work suggests that 254 missense mutations at hotspot positions in KCNA7 may also result in skeletal muscle 255 disorders based on the tissue expression of KCNA7 and the conservation of hotspot positions 256 in the PF00520 domain of this gene.

257

258 Our method also identified six genes with a mutation at the hotspot location that have not 259 previously been associated with NDDs. Three genes - KCNH5, CACNA1B, and KCNG1 -260 have evidence supporting NDD association (Supplementary Table 10). In KCNH5, the same 261 DNM was described as a variant of unknown significance (VUS) in a patient with an epileptic encephalopathy,³¹ and studies are ongoing that associate this gene with NDD in 262 263 other patients (personal communication Prof. Heather Mefford). CACNA1B was recently established as a NDD-associated gene on the basis of LoF DNM enrichment.³² In line with 264 265 this, CACNA1B is the only proposed novel hotspot gene that is predicted to be intolerant to 266 heterozygous loss of function by population constraint (pLI = 1; **Supplementary Data S9**). 267 However, our work suggests that the missense variants we identify at hotspot positions in 268 CACNA1B may be function-altering. KCNG1 has been implicated in neuronal development³³, 269 and the expression profile matches well with that of the other NDD genes that have hotspot 270 mutations. There is also some circumstantial evidence for two of the other three genes. Both 271 TPCN1 and TPCN2 have no prior NDD-association, however, both genes are part of the mTOR complex, which has previously been associated to NDDs.³⁴ Phenotypic data for the 272 273 patient with the missense DNM in TPCN1 shows that this patient has macrocephaly and

severe ASD (**Supplementary Data S14**) which is in line with the fact that mTOR genes have been associated with intracranial volume and intellectual disability.³⁵ However, the only *TPCN1* missense mutation described in literature at a hotspot position (p.102) is associated with early-onset cardiomyopathy.³⁶

278

279 In this analysis, we initially identified stringent mutation hotspots statistically using unique 280 missense mutation counts. While MDHS analysis on even larger cohorts may identify 281 additional hotspots, we believe our method can be used on smaller datasets by also 282 considering recurrent mutations. Applying this lenient method to our cohort identified 32 283 significant missense hotspots (Supplementary Data S3) and no significant hotspots for 284 synonymous or nonsense mutations (Supplementary Data S4-S5). 12 protein domain 285 families had hotspots spanning multiple gene-codons based on 245 DNMs from 67 genes. 48 286 of these 67 genes (72%) are statistically associated with NDDs, representing a 2.53-fold enrichment ($p = 1.26^{-31}$ chi-square test; Supplementary Table 11) and showing the merit of 287 288 this approach. While the inclusion of recurrent mutations allows hotspots driven by 289 proliferative advantages of single mutations in the germline or soma, CpG hypermutability, 290 or biases in clinical ascertainment, it also increases power to detect robust hotspot positions 291 (Supplementary Figure 6). In support of this, we find an enrichment of likely pathogenic 292 missense variants at lenient hotspot positions in clinical databases even if we look only at 293 codons without a mutation in our cohort, showing the merit of this approach. However, 294 additional filtering may be required to remove hotspots driven solely by recurrent missense 295 mutations. Additionally, there is in principle no reason to restrict this method to *de novo* 296 mutations; it could easily be applied to rare inherited variants in large patient cohorts.

297

Kaplanis *et al.* estimate that approximately 350,000 parent-offspring trios would be required to detect the majority of remaining haploinsufficient genes associated with NDDs. Genes with gain-of-function effects are predicted to be even more difficult to detect, even in very large cohorts. Methods based on homology like MDHS provide a novel approach to the identification of disease genes and mechanisms in existing datasets without increasing cohort size.

- 304
- 305

306 Tables							
Gene	Variant	Functional NDD-association	gnomAD AF / SIFT /	DNMs in other	Clinical		
			Polyphen-2 / MPC /	NDD-associated	Interpretation*		
			CADD / MetaDome score	genes	-		
	ENST0000045283		1,20E-02 // Deleterious (0)	2:			
TRPM5	3.1:c.2558G>C;	I la la cara	// Probably damaging (1)	SLC9A1	Uncertain		
<u>*604600</u>	p.R850Q;	Unknown	// - // 28.7 // intolerant	ADNP	(Class 3)		
	PF00520:p.102		(0.49)				
	ENST0000029430		- // Deleterious (0.02) //	0			
TPCN2	9.3:c.1734C>A;	Part of the mTOR complex	Probably damaging				
<u>*612163</u>	p.R545S;	34	(0.965) // 0.80 // 23.5 //				
	PF00520:p.96		slightly intolerant (0.67)				
	ENST0000055078		7.97E-06 // Tolerated (0.1)	0			
TPCN1	5.1:c.963G>A;	Part of the mTOR complex	// Possibly damaging		Likely benign		
<u>*609666</u>	p.R265Q;	34	(0.903) // 2.35 // 26.1 //		(Class 2)		
	PF00520:p.96		tolerant (1.03)				
VCN115	ENST0000032289	Identical VUS (p.327R>H)	- // Deleterious (0) //	0			
*605716	5.7.0.1249G > A; n $P327H$	in unrelated patient with	(0.000) // 1.03 // 32 //				
<u>-003710</u>	p.K32711, $PE00520 \cdot n 102$	epileptic encephalopathy ³¹	(0.999) // 1.93 // 32 //				
	1100520.p.102						
	ENST0000037157		- // Deleterious (0) //	0			
KCNGI	1.4:c.1332G>A;	Involved in neuronal	Probably damaging (1) //				
<u>*603788</u>	<i>p.R349H;</i>	differentiation	2.74 // 32 // highly				
	PF00520:p.102		intolerant (0.13)				
				0			
	E)/(T000000000000000000000000000000000000	Nonsense DNMs in					
GA GIVA ID	ENST000003/13/	CACNA1B lead to a NDD	4,58E-03 // Deleterious (0)				
CACNAIB	2.1 c.188/G>A;	with seizures and	// Probably damaging				
<u>*601012</u>	<i>p.K</i> 58 <i>1H</i> ;	nonepileptic hyperkinetic	(0.999) // 1.32 // 26.1 //				
	PF00520:p.102	movements #618497 ³²	inginy intolerant (0.13)				

307 Table 1. Overview of the novel genes for which we found variants at hotspot positions. First 308 column indicates the gene and OMIM identifier. Second column indicates the identified DNM 309 at the hotspot position. Third column indicates a previously described functional association 310 of the gene or variant to NDD. Fourth column indicate different prediction scores of variant 311 pathogenicity, the fifth indicates if the same patient has any DNMs located in known NDD-312 associated genes, and sixth ACMG classification. Rows marked in grey are novel candidate 313 NDD genes based on additional evidence as described in this paper. Evidence for ACMG 314 classification is provided in Supplementary Table 6. Genomic positions and additional 315 phenotype information for all variants where available can be found in **Supplementary Data** 316 S3 and S14. 317 *Variants were clinically interpreted where proband phenotype information was available 318 (see Supplementary Data S14).

- 319
- 320





325 Figure 1. Workflow of how mutations are extracted from homologous domain regions within 326 genes and aggregated to meta-domain positions. By clockwise orientation, starting in the 327 upper left there are three protein representations of hypothetical genes A, B and C with the 328 mutations identified within a cohort are displayed as red lollipops, the domains as blue and 329 white boxes. The white boxes represent domains that are homologous and are extracted and 330 aligned, including their mutations, and displayed on the upper right part of this image as 331 domains A, B, and C. The mutations within a codon are then aggregated over corresponding 332 homologous domain positions based on sequence alignments to form a meta-domain 333 mutation profile (bottom right). Here, the recurring mutations are counted only once for 334 unique counts (for stringent hotspot identification). The unique counts are the input for 335 variable 'k' to compute a positional MDHS p-value (Equation 1). Together with the total 336 number of mutations 'n' in the meta-domain mutation profile the significance threshold (red 337 dotted line left bottom) can be determined which indicates a meta-domain hotspot if the 338 mutational count exceeds it.

- 339
- 340
- 341
- 342



344 345

Figure 2. The count distribution of missense DNMs aggregated over the Ion Transport protein domain family (PF00520). The total consensus length of this domain is 245 and the sum of the count distribution is 350. The significance threshold is displayed as a dotted black line, computed via the MDHS (Equation 1). The bars that exceeded the significance threshold are coloured in red and represent the mutational hotspots p.96, p.102, and p.231.

353 A.



355 **B.**



357 C.



- 358
- 359
- Figure 3. Changes in structure caused by missense DNMs in NDD-associated genes for each
 hotspot.
- 362 A. Homology model of the KCNQ3 complex with missense DNM p.R227Q marked as a green
- 363 to red change. The KCNQ3 complex is a tetramer constructed from four copies of the
- 364 KCNQ3 monomer. All monomers are marked in different colour shades. This DNM is located
- 365 at identified hotspot p.96. The wild-type Arginine residue is part of the voltage-sensing helix
- 366 and changed into a Glutamine. This change causes it to lose the positive charged that was
- 367 previously found to cause a function-altering mechanism of disease.²⁰

- 368 **B.** Homology model of CACNA1A with missense DNM p.R1663Q marked as a green to red
- 369 change. This DNM is located at identified hotspot p.102. The wild-type Arginine residue is
- 370 part of the voltage-sensing helix and changed into a Glutamine. This change causes it to lose
- 372 *disease*.²¹
- 373 C. Homology model of the KCNH1 complex with missense DNM p.G496R marked as a green
- 374 to red change. The KCNH1 complex is a tetramer constructed from four copies of the
- 375 KCNH1 monomer. All monomers are marked in different colour shades. This DNM is located
- at identified hotspot p.231. The wild-type Glycine residue is near the pore-closing region and
- 377 changed into a much larger Arginine. This may impact pore closure and is previously
- 378 reported to result into a function altering mechanism of disease.²²
- 379



380

381 Figure 4. Hotspot genes are constrained against loss of function and missense variation.

A) *Constraint in hotspot and proposed novel hotspot genes.* The observed variant counts for loss of function (red), missense (orange), and synonymous (pink) variants from the gnomAD v2 release were compared to the expected counts based on a null mutational model (Samocha *et. al.*, 2017). Points represent the mean observed/expected ratios for all genes in each set and bars denote the mean upper and lower bound fractions for these ratios. The dashed line at observed/expected = 1 indicates perfect adherence to the null mutational model (observed set); values that fall below this line are constrained.

B) Mutation hotspots occur in missense constrained regions within genes. Regional missense

constraint was compared across PF00520 domain-containing control genes (blue), PF00520
 domain-containing NDD-associated genes (green) and hotspot genes (yellow).



393

Figure 5. Hotspot genes have a distinct gene expression pattern.

A) *Tissue expression of hotspot genes compared to control and other NDD genes*. Expression of the 19 established NDD genes containing missense DNM(s) at a stringent mutation hotspots (hotspot genes, yellow) were evaluated across 54 GTEx tissues (x-axis). Hotspot gene expression was compared to NDD genes (green) and control genes (blue). The y-axis depicts the proportion of expressed genes (Methods). Squares and bars depict the median and standard deviation respectively of NDD and control gene distributions.

401 **B**) *TPM distribution in brain and other tissues varies across gene sets.* Control (blue) and 402 NDD-associated (green) genes are represented by 2D density distributions. Hotspot genes

403 (yellow) are shown as points, as are NDD-associated genes containing a PF00520 domain404 (grey). Proposed novel hotspot genes are marked by their gene name in red text.



406 407

408 Figure 6. Lenient hotspots are enriched for likely pathogenic missense variation in 409 clinical databases. Counts of likely pathogenic (LP, red), uncertain (VUS, grey) and likely 410 benign (LB, blue) missense variants in VKGL (A) and ClinVar (B) in domains containing a 411 lenient hotspot position. The proportion of LP missense variants (LP/(LP + LB), see 412 Methods) was compared between mutation hotspots (purple) and all other protein consensus 413 positions within the domain (orange). This comparison was done for all possible missense 414 variants in these domains (row 2) and with positions containing DNMs in our cohort 415 excluded (row 3).



417

Figure 7. Patients are enriched for missense variants in lenient hotspot positions compared to healthy population controls. A) NDD and ASD are enriched for missense DNMs in lenient hotspot positions compared to unaffected individuals (green) but are not enriched for synonymous DNMs (grey). Only DNMs within protein consensus positions were used for this comparison (see Methods). B) ASD probands are enriched for missense DNMs in lenient hotspot positions (green) not present in our NDD cohort.

425 Materials and Methods

426 **Dataset of** *de novo* **mutations**

We obtained the set of 45,221 DNMs from the Kaplanis *et al* study.⁷ These DNMs were identified in 31,058 DD patients combined across three centres. The genetic testing approach of these patients were described previously per centre: DDD,³ GeneDX,³⁷ and, Radboudumc.⁴ All individuals that underwent genetic testing provided informed consent.⁷ Subsets of these patients have been analysed and reported in previous publications.^{3,23,37,38}

432

433 Developmental disorder diagnostic gene lists

We use the diagnostic lists of DD-associated genes from the Kaplanis *et al.* study. We consider all genes statistically associated to NDDs in this study to be NDD genes (novel, consensus, and discordant genes, total genes = 1010).⁷

437

Additionally, we used the Deciphering Developmental Disorders Genotype2Phenotype (DDG2P, accessed 22-04-2021) list to assess the burden of gain of function mutational mechanisms already described in hotspot gene families. We considered activating, gain of function, dominant negative, and increased gene dosage mutation consequences in this list to be gain of function.

443

444 Annotation of transcript details, protein and meta-domain position annotation

- The DNMs were annotated with corresponding GENCODE³⁹ transcripts from release 19 445 446 GRCh37.p13 Basic set, protein information from UniProtKB/Swiss-Prot⁴⁰ Release 2016_09, Pfam-A⁴¹ v30.0 protein domains information, and meta-domain¹⁰ positions using a local 447 MetaDome¹⁴ 448 version of the web available server (code at 449 https://github.com/cmbi/metadome). Meta-domains are multiple sequence alignments of 450 regions within human protein-coding genes that correspond to Pfam protein domain families. 451 The DNMs that correspond to Pfam consensus positions are annotated with the corresponding 452 Pfam domain ID and consensus position.
- 452 Fram domain iD and consensus 453

454 Filtering the annotated DNMs

- The annotation process can result in multiple GENCODE gene transcripts per DNM. To ensure a single GENCODE transcript per gene we performed a filtering step by the following order of criteria:
- 458458459459a Filter to variants that with transcript consequence: missense, synonymous, or, stopgained
- 460 2. The transcript corresponds to a human canonical or isoform entry in Swiss-Prot
- 461
 461
 462
 3. This transcript contains all (or most) of the *de novo* mutations for the corresponding gene
- 463 4. The transcript translates to the longest protein sequence length
- 464 5. If multiple transcripts remain for a gene, one of these is selected
- 465 6. Filter variants only to those that are in a Pfam protein domain

467 MDHS: Detection of variant hotspots in homologous protein domains

The Pfam domain ID and consensus position allows for aggregation of genetic variants
through meta-domain positions. To identify meta-domain positions that are significantly
enriched with variants we created the MDHS (Meta-Domain HotSpot) p-value as follows:

MDHS p-value =
$$Pr\left(x < k; Bin\left(n, \frac{1}{L}\right)\right)$$
 (1.)

472

473 In the context of meta-domains, n corresponds to the total number of aggregated genetic 474 variants for the Pfam domain ID, L is the total number of possible consensus positions for a 475 Pfam domain ID, k is the total number of genetic variants aggregated at a single consensus 476 position, and x = k - 1, which depicts the chance of finding less than observed genetic 477 variants at the consensus position. The MDHS p-value is adapted from the mCluster¹⁶ and 478 DS-Score¹⁷. In line with these methods, variants are assumed to follow a Binomial 479 distribution. We correct the MDHS p-value via the Bonferroni method for the total number of 480 Pfam protein domain IDs considered. If a Bonferroni corrected MDHS p-value < 0.05, we 481 consider it as a significant mutational hotspot.

482

Our code to analyse the MDHS p-value was optimized to only compute for domains which can have significant hotspots. It implements a filter for domain families which works as follows: 1.) Count the number of domain consensus positions with one or more variant as 'n_hotspot_candidates'. 2.) Count the number of DNMs that span at least more than one unique protein position at each consensus position and sum them up to represent 'n_hotspots_with_variation_from_more_then_one_protein_position'. 3.) Apply the filter criteria such that each domain family abides:

490

$\frac{n_hotspots_with_variation_from_more_then_one_protein_position}{n_hotspot_candidates} > 1$

491

492 See the value at column 'hotspot_uniqueness' in **Supplementary Data S2, S5, S6**.

493

494 Stringent and lenient counting of variants in MDHS

495 We use two ways to determine variable k in the MDHS p-value (**Equation 1**). Unless 496 otherwise specified, we count unique variants by considering mutated chromosomal positions 497 only once, thereby reducing the impact of recurrent mutations in a single gene (stringent). 498 Alternatively, we refer to a lenient way of variant counting when we count every mutation 499 equally (including recurrent mutations). For a schematic, see **Figure 1**.

500

501 **Functional characterization**

We used the Ensembl Variant Effect Predictor (VEP)⁴² to annotate gnomAD allele frequency 503 (AF)²⁴, SIFT⁴³, Polyphen-2⁴⁴, MPC²⁵, and the CADD_Phred⁴⁵. MetaDome¹⁴ tolerance 504 indication based on regional d_N/d_S was obtained manually. ACMG⁴⁶ classification was 505 obtained through variant curation by a laboratory specialist. Available phenotype information 506 for patients with missense mutations in hotspot positions can be found in **Supplementary** 507 **Data S14**.

508

509 **Protein 3D structure modelling of the genes with identified hotspots**

For each of the 25 genes with a DNM located at one of the hotspots we submitted the corresponding protein sequence (based on the transcript of **Filtering the annotated DNMs**) to the YASARA & WHAT IF Twinset^{47,48} homology modelling script using the default settings. The regions corresponding to the PF00520 were extracted from all resulting homology structures and combined in a single YASARA scene and then structurally aligned using the MOTIF script. The structures are available in **Supplementary Data S7**. The protein structure effects of mutations have been reported in **Supplementary Data S8**.

518 **Population constraint**

519 Constraint information, including observed and expected counts, z-scores, and pLI, pNull, $\frac{1}{24}$

520 and pRec were calculated on gnomAD v2.1.1. 24

521

522 Regional missense constraint

523 Genes with regions of differential missense constraint were identified as described by 524 Samocha *et. al*, 2017 in the $ExAC^{49}$ dataset. In brief, the fraction of observed missense 525 variation along a transcript was tested for uniformity using a likelihood ratio test. If the 526 distribution was not uniform, the transcript was considered to have evidence of regional 527 missense constraint. 2,700 genes showed evidence of at least two regions of distinct missense 528 constraint using this method.

529

530 Expression data

531 Pre-computed tissue expression values in transcripts per million (TPM) were taken from 532 GTEx v8 (GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_median_tpm.gct.gz).^{26,50}

533

534 Gene sets for constraint and expression analysis

535 The set of 56,200 genes for which median TPM values were available was divided into one 536 of four sets: proposed novel hotspot genes, hotspot genes, NDD-associated genes, and control 537 genes. Genes containing a mutation hotspot were divided into two categories: hotspot genes, 538 n = 19, and proposed novel hotspot genes, n = 6. These categories were distinguished by 539 presence on the DD gene list (see **Developmental disorder diagnostic gene lists**); hotspot 540 genes are present on this list while proposed novel hotspot genes are not. The remaining 541 genes were divided into NDD-associated genes (n = 992, excluding hotspot genes) and 542 control genes not statistically associated with intellectual disability or developmental delay (n 543 = 55,183). For some analyses, control genes present in DDG2P (n = 1,250, accessed 22-04-544 2021) or OMIM (n = 3,402, accessed 31-08-2021) but not statistically associated to NDDs 545 were considered separate classes. Additionally, only some of these genes had population 546 constraint information available from gnomAD v2.1.1 (n = 19,658). Genes in all sets can be 547 found in Supplementary Data S15.

548

549 Of the 56,200 genes described above, 105 contained a PF00520 domain in MetaDome v1.0.1. 550 These 105 genes represented 19 hotspot genes, 6 proposed novel hotspot genes, 12 NDD-551 associated genes not containing a missense DNM at a hotspot position in our cohort, and 68 552 control genes (of which 6 are present in DDG2P and 26 in OMIM; **Supplementary Data** 553 **S16**).

554

555 **Proportion of expressed genes across GTEx tissues**

556 A fixed level of TPM > 1 was used to define expression in each tissue. NDD-associated and 557 control genes were randomly sampled 1000 times into sets of 19 genes, and the proportion of 558 expressed genes (number of genes with TPM > 1/total number of genes) was calculated for 559 each set. This generated a distribution of proportions across 54 GTEx tissues. The proportion 560 of expressed hotspot genes per tissue was computed without sampling.

561

562 **TPM differences between tissue groups**

563 In order to assess expression differences between brain and other tissues, GTEx tissues were 564 divided into two groups. We considered the amygdala, anterior cingulate cortex (BA24), 565 caudate (basal ganglia), cerebellar hemisphere, cerebellum, cortex, frontal cortex (BA9), 566 hippocampus, hypothalamus, nucleus accumbens (basal ganglia), putamen (basal ganglia), 567 spinal cord (cervical c-1), and substantia nigra brain tissues (n = 12). All non-brain tissues 568 were included in the 'other tissues' set (n = 42). The TPM value for each tissue set was 569 defined as the median TPM of all tissues in the set. Based on these differences, we modelled 570 the brain TPM distribution in hotspot genes and control genes and the other tissue TPM

distribution in hotspot genes and NDD-associated genes as normal distributions. For each set
of distributions, a likelihood ratio of belonging to each distribution was calculated. Proposed
novel hotspot genes were considered to have evidence for association with NDDs if they
were more likely to belong to the hotspot gene distribution across both tests.

575

576 Filtering and annotation of additional *de novo* mutation cohorts

577 We analysed the enrichment of missense and synonymous DNMs in lenient hotspot positions 578 across a total of three additional published DNM cohorts. We used an autism-spectrum 579 disorder (ASD) cohort published by Satterstrom et al. (35,584 total individuals, 11,986 with 580 ASD), a congenital heart defect (CHD) cohort published by Jin et al. (2,645 trios), and a 581 cohort of healthy individuals sequenced by Jonsson et al. (1,548 trios). To increase our power 582 to find significant differences at lenient hotspot positions, we pooled the Jonsson et al. 583 healthy individuals with unaffected siblings from the Satterstrom ASD cohort (1,740 siblings) 584 for a total of 3,288 unaffected individuals.

585

Annotation of meta-domain protein consensus positions for these datasets was done as
 previously described for Kaplanis *et al.* using MetaDome. The number of PTV, missense, and
 synonymous SNVs in Pfam protein domains in each dataset can be found in Supplementary
 Table 12.

590

591 Variant annotation

Variants at protein consensus positions were checked for clinical interpretation across four
curated variant databases: ClinVar, HGMDPro, Swiss-Prot and VKGL, all accessed 21-082021. Mapping of protein consensus positions to GRCh37 genomic positions for each gene as
done using MetaDome.

For analysis on stringent hotspot positions, ClinVar data was unfiltered on evidence level or
review status. We classified missense variation as LP (pathogenic or likely pathogenic,
ACMG Class V or IV) or VUS (variant of uncertain significance, ACMG Class III) based on
the most severe class across all four databases (Supplementary Data S13).

600 For the enrichment analysis on lenient hotspot positions, only ClinVar and VKGL data was 601 used because these two databases include likely benign variants. ClinVar data was filtered on 602 review status (required to be one of 'practice guideline', 'reviewed by expert panel', 'criteria 603 provided, multiple submitters, no conflicts', 'criteria provided, single submitter'). Variants 604 were classified as LP (ACMG Class V or IV), VUS (ACMG Class III) or LB (benign or 605 likely benign, ACMG Class II or I). Variants with conflicting interpretations of pathogenicity 606 within or between databases (LP and LB annotations) were removed, as were variants with 607 only VUS annotations.

608

609 Code availability and data access

610 Constraint and expression analysis was done in R version 3.6.2. Code and data to reproduce available 611 all parts of the analysis is on GitHub 612 (https://github.com/cmbi/MetaDomainHotSpot). MetaDome was used to annotate meta-613 domains. for the source code this is also available Github on 614 (https://github.com/cmbi/metadome).

616 Acknowledgements

617 We thank Dr. Torti and Dr. Retterer from GeneDX for connecting us with Prof. Mefford. We

618 thank Prof. Mefford for information regarding the likely NDD-association of KCHN5. We

- 619 thank Elke de Boer for useful discussions. This work was in part financially supported by
- 620 grants from the Netherlands Organization for Scientific Research (916-14-043 to C.G. and
- 621 918-15-667 to J.A.V.), and from the Radboud Institute for Molecular Life Sciences, Radboud
- 622 University Medical Center (R0002793 to G.V.).

623 **URLs**

- 624 YASARA: <u>http://www.yasara.org/</u>
- 625 CATH-Gene3D: <u>http://www.cathdb.info/</u>
- 626 Variant Effect Predictor: <u>https://www.ensembl.org/Tools/VEP</u>
- 627 GTEx: <u>https://www.gtexportal.org/home/</u>
- 628 MetaDome web server: <u>https://stuart.radboudumc.nl/metadome/</u>
- 629 MetaDome GitHub repository: <u>https://github.com/cmbi/metadome</u>
- 630 MetaDomainHotspot analyses repository: https://github.com/cmbi/MetaDomainHotSpot
- 631 DECIPHER: https://www.deciphergenomics.org/
- 632 HGMD: <u>http://www.hgmd.cf.ac.uk/ac/index.php</u>
- 633 ClinVar: <u>https://www.ncbi.nlm.nih.gov/clinvar/</u>
- 634 Swiss-Prot: <u>https://www.uniprot.org/</u>
- 635 VKGL: https://www.vkgl.nl/nl/diagnostiek/vkgl-datashare-database

636 medRxiv Note

- 637 Our **Supplementary Data S7** file (YASARA protein structures) cannot be made available
- 638 through medRxiv. Please contact the corresponding author for this data.
- 639

References 640 641 1. Veltman, J. a & Brunner, H. G. De novo mutations in human genetic disease. Nat. Rev. 642 Genet. 13, 565–75 (2012). 643 2. Martin, H. C. et al. Quantifying the contribution of recessive coding variation to 644 developmental disorders. Science (80-.). 362, 1161–1164 (2018). 645 3. Deciphering Developmental Disorders Study. Prevalence and architecture of de novo 646 mutations in developmental disorders. *Nature* 542, 433–438 (2017). 647 de Ligt, J. et al. Diagnostic exome sequencing in persons with severe intellectual 4. 648 disability. N. Engl. J. Med. 367, 1921-9 (2012). 649 5. Deciphering Developmental Disorders Study et al. Large-scale discovery of novel 650 genetic causes of developmental disorders. *Nature* **519**, 223–228 (2015). 651 6. Turner, T. N. et al. denovo-db: a compendium of human de novo variants. Nucleic 652 Acids Res. 45, D804–D811 (2017). 653 7. Kaplanis, J. et al. Evidence for 28 genetic disorders discovered by combining 654 healthcare and research data. Nature (2020). doi:10.1038/s41586-020-2832-5 655 8. Geisheker, M. R. et al. Hotspots of missense mutation identify neurodevelopmental 656 disorder genes and functional domains. *Nat. Neurosci.* **20**, 1043–1051 (2017). 9. 657 Lelieveld, S. H. et al. Spatial Clustering of de Novo Missense Mutations Identifies 658 Candidate Neurodevelopmental Disorder-Associated Genes. Am. J. Hum. Genet. 101, 478-484 (2017). 659 660 10. Wiel, L., Venselaar, H., Veltman, J. A., Vriend, G. & Gilissen, C. Aggregation of 661 population-based genetic variation over protein domain homologues and its potential 662 use in genetic diagnostics. Hum. Mutat. 1-10 (2017). doi:10.1002/humu.23313 663 11. Stenson, P. D. et al. The Human Gene Mutation Database: towards a comprehensive 664 repository of inherited mutation data for medical research, genetic diagnosis and next-665 generation sequencing studies. Hum. Genet. 136, 1-13 (2017). 666 12. Landrum, M. J. et al. ClinVar: public archive of interpretations of clinically relevant 667 variants. Nucleic Acids Res. 44, D862-8 (2016). 668 13. Schuurs-Hoeijmakers, J. H. M. et al. Recurrent de novo mutations in PACS1 cause 669 defective cranial-neural-crest migration and define a recognizable intellectual-670 disability syndrome. Am. J. Hum. Genet. 91, 1122–7 (2012). 671 14. Wiel, L. et al. MetaDome: Pathogenicity analysis of genetic variants through 672 aggregation of homologous human protein domains. Hum. Mutat. 40, humu.23798 673 (2019). 674 15. Peterson, T. A., Park, D. H. & Kann, M. G. A protein domain-centric approach for the 675 comparative analysis of human and yeast phenotypically relevant mutations. BMC 676 Genomics 14 Suppl 3, (2013). 677 16. Yue, P. et al. Inferring the functional effects of mutation through clusters of mutations 678 in homologous proteins. Hum. Mutat. 31, 264–271 (2010). 679 17. Peterson, T. A., Nehrt, N. L., Park, D. H. & Kann, M. G. Incorporating molecular and 680 functional context into the analysis and prioritization of human variants associated 681 with cancer. J. Am. Med. Informatics Assoc. 19, 275–283 (2012).

- Sillitoe, I., Lewis, T. & Orengo, C. Using CATH-Gene3D to Analyze the Sequence,
 Structure, and Function of Proteins. *Curr. Protoc. Bioinforma.* 50, 1.28.1-1.28.21
 (2015).
- 685 19. Bezanilla, F. How membrane proteins sense voltage. *Nat. Rev. Mol. Cell Biol.* 9, 323–332 (2008).
- Sands, T. T. *et al.* Autism and developmental disability caused by KCNQ3 gain-of-function variants. *Ann. Neurol.* 86, 181–192 (2019).
- 689 21. Luo, X. et al. Clinically severe CACNA1A alleles affect synaptic function and

690		neurodegeneration differentially. <i>PLoS Genet.</i> 13 , e1006905 (2017).
691	22.	Kortüm, F. <i>et al.</i> Mutations in KCNH1 and ATP6V1B2 cause Zimmermann-Laband
692		syndrome. Nat. Genet. 47, 661–667 (2015).
693	23.	Lelieveld, S. H. <i>et al.</i> Meta-analysis of 2,104 trios provides support for 10 new genes
694		for intellectual disability. <i>Nat. Neurosci.</i> 19 , 1194–1196 (2016).
695	24.	Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation
696		in 141.456 humans. <i>Nature</i> 581 , 434–443 (2020).
697	25.	Samocha, K. E. et al. Regional missense constraint improves variant deleteriousness
698		prediction. <i>bioRxiv</i> 148353 (2017). doi:10.1101/148353
699	26.	Aguet, F. et al. The GTEx Consortium atlas of genetic regulatory effects across human
700		tissues. Science (80). 369 , 1318–1330 (2020).
701	27.	Daniil, G. et al. CACNA1H Mutations Are Associated With Different Forms of
702		Primary Aldosteronism. EBioMedicine 13, 225–236 (2016).
703	28.	Zhang, X. Y. <i>et al.</i> Gain-of-function mutations in SCN11A cause familial episodic
704		pain. Am. J. Hum. Genet. 93, 957–66 (2013).
705	29.	Heyne, H. O. <i>et al.</i> De novo variants in neurodevelopmental disorders with epilepsy.
706		Nat. Genet. 50, 1048–1053 (2018).
707	30.	Heyne, H. O. et al. Predicting functional effects of missense variants in voltage-gated
708		sodium and calcium channels. Sci. Transl. Med. 12, eaay6848 (2020).
709	31.	Veeramah, K. R. et al. Exome sequencing reveals new causal mutations in children
710		with epileptic encephalopathies. Epilepsia 54, 1270–1281 (2013).
711	32.	Gorman, K. M. et al. Bi-allelic Loss-of-Function CACNA1B Mutations in Progressive
712		Epilepsy-Dyskinesia. Am. J. Hum. Genet. 104, 948–956 (2019).
713	33.	Chiocchetti, A. G. et al. Transcriptomic signatures of neuronal differentiation and their
714		association with risk genes for autism spectrum and related neuropsychiatric disorders.
715		Transl. Psychiatry 6, (2016).
716	34.	Cang, C. et al. mTOR regulates lysosomal ATP-sensitive two-pore Na(+) channels to
717		adapt to metabolic state. Cell 152, 778–790 (2013).
718	35.	Reijnders, M. R. F. et al. Variation in a range of mTOR-related genes associates with
719		intracranial volume and intellectual disability. Nat. Commun. 8, 1052 (2017).
720	36.	Reuter, M. S. et al. The Cardiac Genome Clinic: implementing genome sequencing in
721		pediatric heart disease. Genet. Med. 22, 1015–1024 (2020).
722	37.	Retterer, K. et al. Clinical application of whole-exome sequencing across clinical
723		indications. Genet. Med. 18, 696–704 (2016).
724	38.	Wright, C. F. et al. Genetic diagnosis of developmental disorders in the DDD study: a
725		scalable analysis of genome-wide research data. Lancet 385, 1305–1314 (2015).
726	39.	Harrow, J. et al. GENCODE: The reference human genome annotation for The
727		ENCODE Project. Genome Res. 22, 1760–1774 (2012).
728	40.	Boutet, E. et al. UniProtKB/Swiss-Prot, the Manually Annotated Section of the
729		UniProt KnowledgeBase: How to Use the Entry View. Methods Mol. Biol. 1374, 23-
730		54 (2016).
731	41.	Finn, R. D. et al. The Pfam protein families database: towards a more sustainable
732		future. Nucleic Acids Res. 44, D279–D285 (2016).
733	42.	McLaren, W. et al. The Ensembl Variant Effect Predictor. Genome Biol. 17, 1–14
734		(2016).
735	43.	Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous
736		variants on protein function using the SIFT algorithm. Nat. Protoc. 4, 1073–1081
737		(2009).
738	44.	Adzhubei, I. A. et al. A method and server for predicting damaging missense
739		mutations. <i>Nat. Methods</i> 7 , 248–9 (2010).

- Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* 46, 310–315 (2014).
- Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants:
 a joint consensus recommendation of the American College of Medical Genetics and
 Genomics and the Association for Molecular Pathology. *Genet. Med.* 17, 405–24
 (2015).
- Krieger, E., Koraimann, G. & Vriend, G. Increasing the precision of comparative models with YASARA NOVA—a self-parameterizing force field. *PROTEINS Struct. Funct. Bioinforma.* 47, 393–402 (2002).
- Vriend, G. WHAT IF: A molecular modeling and drug design program. *J. Mol. Graph.*50 8, 52–56 (1990).
- 49. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature*536, 285–291 (2016).
- Aguet, F. *et al.* Genetic effects on gene expression across human tissues. *Nature* 550, 204–213 (2017).
- 755 756