

1 **Genetic constraint at single amino acid resolution improves missense variant**
2 **prioritisation and gene discovery**

3

4 Xiaolei Zhang^{1,2,3}, Pantazis I. Theotakis^{1,2,3}, Nicholas Li^{1,2,3}, the SHaRe Investigators,
5 Caroline F. Wright⁴, Kaitlin E. Samocha^{5,6}, Nicola Whiffin^{7*#}, James S. Ware^{1,2,3*#}

6

7 1. National Heart & Lung Institute, Imperial College London, London W12 0NN,
8 United Kingdom

9 2. MRC London Institute of Medical Sciences, Imperial College London, London
10 W12 0NN, United Kingdom

11 3. Royal Brompton & Harefield Hospitals, Guy's and St. Thomas' NHS Foundation
12 Trust, London SW3 6NP, United Kingdom

13 4. University of Exeter Medical School, Institute of Biomedical and Clinical Science,
14 Royal Devon and Exeter Hospital, Exeter EX2 5DW, UK

15 5. Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA
16 02114, USA

17 6. Program in Medical and Population Genetics, Broad Institute of MIT and Harvard,
18 Cambridge, MA 02142, USA

19 7. Wellcome Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK

20 *These authors contributed equally

21 #To whom correspondence should be addressed: nwhiffin@well.ox.ac.uk or

22 j.ware@imperial.ac.uk

23

24

25

26 **the SHaRe Investigators**

27 Euan Ashley¹, Steven D. Colan², Sharlene M. Day³, Adam Helms⁴, Carolyn Y. Ho⁵,
28 Jodie Ingles^{6,7,8}, Daniel Jacoby⁹, Neal K Lakdawala⁵, Michelle Michels¹⁰, Iacopo
29 Olivotto¹¹, Anjali Owens³, Victoria N Parikh¹, Alexandre C. Pereira^{1,2}, Joseph
30 Rossano¹³, Sara Saberi⁴, Chris Semsarian^{8,14,15}, Samuel Wittekind¹⁶.

31 1. Division of Cardiovascular Medicine, Stanford University Medical Center, Stanford,
32 CA, USA;

33 2. Department of Cardiology, Boston Children's Hospital, Boston, MA, USA;

34 3. Division of Cardiovascular Medicine and Penn Cardiovascular Institute, Perelman
35 School of Medicine, University of Pennsylvania, Philadelphia, USA;

36 4. Department of Internal Medicine-Cardiology, University of Michigan, Ann Arbor,
37 MI, USA.

38 5. Cardiovascular Division, Brigham and Women's Hospital, Boston, MA, USA;

39 6. Centre for Population Genomics, Garvan Institute of Medical Research, and
40 UNSW Sydney, Sydney, Australia

41 7. Centenary Institute, The University of Sydney, Sydney, Australia;

42 8. Department of Cardiology, Royal Prince Alfred Hospital, Sydney, Australia;

43 9. Department of Internal Medicine, Yale University, New Haven, CT, USA;

44 10. Department of Cardiology, Thoraxcenter, Erasmus MC Rotterdam, Rotterdam,
45 Netherlands;

- 46 11. Cardiomyopathy Unit, Careggi University Hospital, Florence, Italy;
- 47 12. Heart Institute (InCor), University of Sao Paulo Medical School, Sao Paulo,
48 Brazil;
- 49 13. Children's Hospital of Philadelphia, PA, USA
- 50 14. Agnes Ginges Centre for Molecular Cardiology Centenary Institute, The
51 University of Sydney, Australia;
- 52 15. Sydney Medical School, Faculty of Medicine and Health, The University of
53 Sydney, Australia;
- 54 16. Cincinnati Children's Hospital Medical Center, Heart Institute, Cincinnati, OH,
55 USA
- 56
- 57
- 58
- 59
- 60
- 61
- 62
- 63
- 64
- 65
- 66
- 67
- 68

69 **Abstract**

70 The clinical impact of most germline missense variants in humans remains unknown.
71 Genetic constraint identifies genomic regions under negative selection, where
72 variations likely have functional impacts, but the spatial resolution of existing
73 constraint metrics is limited. Here we present the Homologous Missense Constraint
74 (HMC) score, which measures genetic constraint at quasi single amino-acid
75 resolution by aggregating signals across protein homologues. We identify one million
76 possible missense variants under strong negative selection. HMC precisely
77 distinguishes pathogenic variants from benign variants for both early-onset and
78 adult-onset disorders. It outperforms existing constraint metrics and pathogenicity
79 meta-predictors in prioritising *de novo* mutations from probands with developmental
80 disorders (DD), and is orthogonal to these, adding power when used in combination.
81 We demonstrate utility for gene discovery by identifying seven genes newly-
82 significant associated with DD that could act through an altered-function mechanism.
83 Overall, HMC is a novel and strong predictor to improve missense variant
84 interpretation.

85

86 **Main**

87 Quantifying the depletion of natural variation in human populations provides a
88 powerful approach to identify variants of large effect¹⁻⁸. Since variants causing
89 severe early-onset disorders are under selective pressure, they are observed less
90 often than functionally neutral variants. Such depletion of genetic variation
91 (constraint) has been shown to provide strong evidence to prioritise disease-
92 associated genes¹⁻³, identify critical regions within genes^{4,5}, and investigate the
93 effect of non-coding variants⁶⁻⁸.

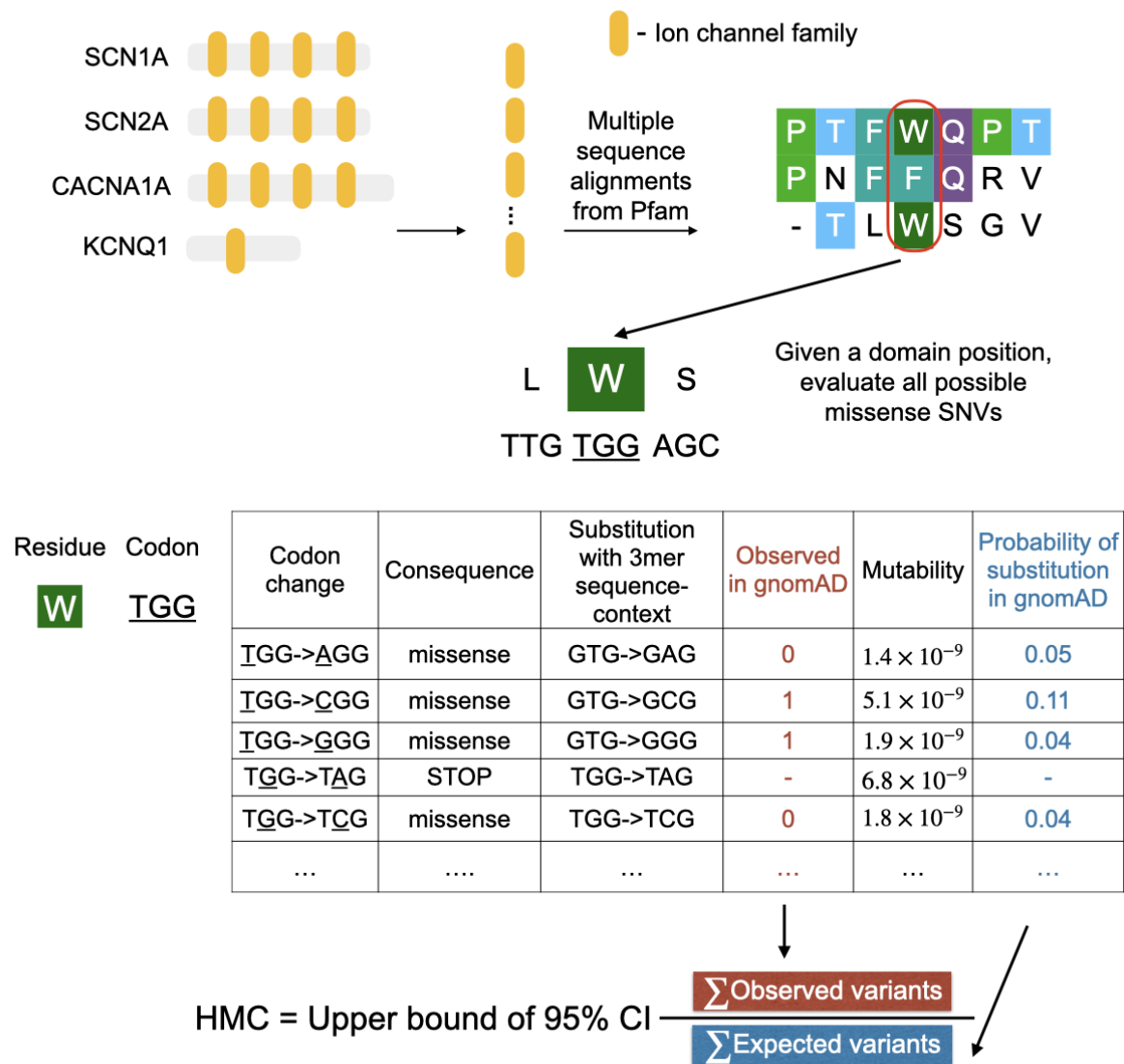
94

95 However, these existing metrics have limited resolution to analyse individual
96 residues, and limited application in genes with sparsely distributed pathogenic
97 missense variants since they explicitly rely on signals clustered linearly within
98 genes¹⁻⁵ (**Figure S1**). To address this issue, we sought to develop an amino-acid
99 level constraint metric. Since we are still underpowered to evaluate the depletion of
100 variants at individual residues, we evaluate homologous residues instead. While
101 previous studies have aggregated variant information over homologous residues to
102 infer functional effect⁹⁻¹⁴, a pathogenicity predictor using genetic constraint hasn't
103 been studied.

104

105 Here we developed an amino-acid level constraint metric by aggregating the signal
106 over evolutionarily equivalent positions across human protein domains. While there
107 are alternative definitions of homology, we use protein domain families defined by
108 the Pfam database¹⁵, which identifies regions of homology in most genes (see
109 **Supplementary Information** for discussion on alternative approaches). Of 70
110 million possible missense variants (defined by NCBI RefSeq Select transcripts¹⁶) in
111 the human genome, 28 million are mapped to Pfam protein domain families. After
112 excluding residues with limited statistical power (see **Method**), 15 million rare
113 missense variants (~21% of all possible missense variants) are assessable, which
114 are defined as those with a minor allele frequency (MAF) < 0.1 or absent from the
115 Genome Aggregation Database (gnomAD v2.1.1; 125,748 samples)². Given a set of
116 homologous residues, we calculated the genetic intolerance of missense variants at
117 individual residues as the ratio of the number of rare missense variants observed in
118 the 125K gnomAD population (Observed) to the number of neutral substitutions

119 expected (Expected) given tri-nucleotide sequence context, CpG methylation levels,
120 and sequencing coverage, using a null model described previously² (**Figure 1**). The
121 Homologous Missense Constraint (HMC) score is defined as the upper bound of the
122 95% CI of the Observed/Expected ratio. A protein residue with HMC score <1
123 indicates that missense variants affecting the homologous residues are significantly
124 under negative selection (P -value < 0.05) and likely to be deleterious. 3,304,332
125 possible missense variants (21.7% of assessable) overlie constrained residues with
126 HMC<1. We further classified 1,322,835 possible missense variants (8% of
127 assessable) at highly constrained residues, using a more stringent threshold of
128 HMC<0.8, which we find clinically relevant, as demonstrated below.
129



130

131 **Figure 1. Overview of developing Homologous Missense Constraint.** Here we
 132 illustrate how to calculate HMC scores using a subset of genes with ion channel
 133 domains (Pfam ID: PF00520). Evolutionary equivalent residues were identified by
 134 aligning protein sequences across the protein family. Given a domain position with
 135 equivalent residues, the Observed/Expected ratio is calculated to measure the
 136 genetic constraint of missense variants at this domain position. HMC score is defined
 137 as the upper bound of 95% CI of the Observed/Expected ratio. Missense variants
 138 affecting domain positions with $HMC < 1$ are significantly ($P\text{-value} < 0.05$) under
 139 genetic intolerance thus predicted as deleterious.

140

141 We robustly evaluated the utility of HMC scores using a wide range of independent
142 tasks, including (i) assessing classification performance of HMC benchmarked
143 against “gold-standard” variant interpretations from ClinVar; (ii) assessing whether
144 HMC prioritises disease-associated variants using case-control analyses in cohorts
145 of patients with known disease phenotypes, without reliance on a gold-standard
146 variant interpretation as a reference; and (iii) evaluating HMC against reference
147 variants evaluated using *in vitro* assays of altered function.

148

149 First, we showed that HMC can distinguish pathogenic variants from benign variants
150 in ClinVar. We found that ClinVar pathogenic variants are significantly enriched at
151 constrained domain positions ($HMC < 1$; $Rate_{Pathogenic\ vs\ Benign} = 3.9$, $95\%CI = 3.5-4.2$, P -
152 $value = 1 \times 10^{-304}$) and significantly depleted at unconstrained domain positions
153 ($HMC \geq 1$; $Rate_{Pathogenic\ vs\ Benign} = 0.62$, $95\%CI = 0.61-0.64$, P -value = 1×10^{-311}) (**Figure**
154 **2a**). The strength of the association increases as domain residues are under
155 stronger genetic constraint ($HMC < 0.5$; $Rate_{Pathogenic\ vs\ Benign} = 37.9$, $95\%CI = 15.7-91.3$,
156 P -value = 1×10^{-41}) indicating that variants with lower HMC scores are more likely to
157 be disease-causing.

158

159 Next, we asked whether HMC could prioritise deleterious *de novo* mutations (DNMs).
160 We analysed published DNMs identified in 5,264 probands ascertained with severe
161 neurodevelopmental delay (NDD) and 2,179 unaffected controls¹⁷. We found that *de*
162 *novo* missense mutations in highly constrained domain positions ($HMC < 0.8$) are
163 significantly enriched in NDD cases ($Rate_{NDD\ cases\ vs\ controls} = 4.1$, $95\% CI = 2.4-6.9$, P -
164 $value = 3.1 \times 10^{-10}$; **Figure 2b-c**). Similarly, highly constrained DNMs are significantly
165 enriched in probands ascertained with autism spectrum disorders ($Rate_{ASD\ cases\ vs$

166 controls=2.2, 95% CI=1.3-3.7, P -value=0.0028; **Figure S4**). In a larger trio cohort with
167 31,058 probands of developmental disorders (referred hereafter as “the 31K DD
168 cohort”)¹⁸, we further evaluated the enrichment of constrained DNMs (the ratio of
169 observed to background expectation^{19,20}) in 285 dominant DD-associated genes that
170 showed statistical enrichment of DNMs in that cohort. While missense variants
171 located in annotated domains have a higher burden than those located elsewhere
172 (Obs/Exp=13.6, 95%CI=12.9-14.3), HMC can further narrow down to a subset as
173 highly constrained (<0.8) with an effect size close to that of protein-truncating
174 variants (Obs/Exp=27.6, 95%CI=25.5-29.7 vs PTV: Obs/Exp=32.4, 95%CI=30.1-
175 34.0; **Figure 2d**).

176

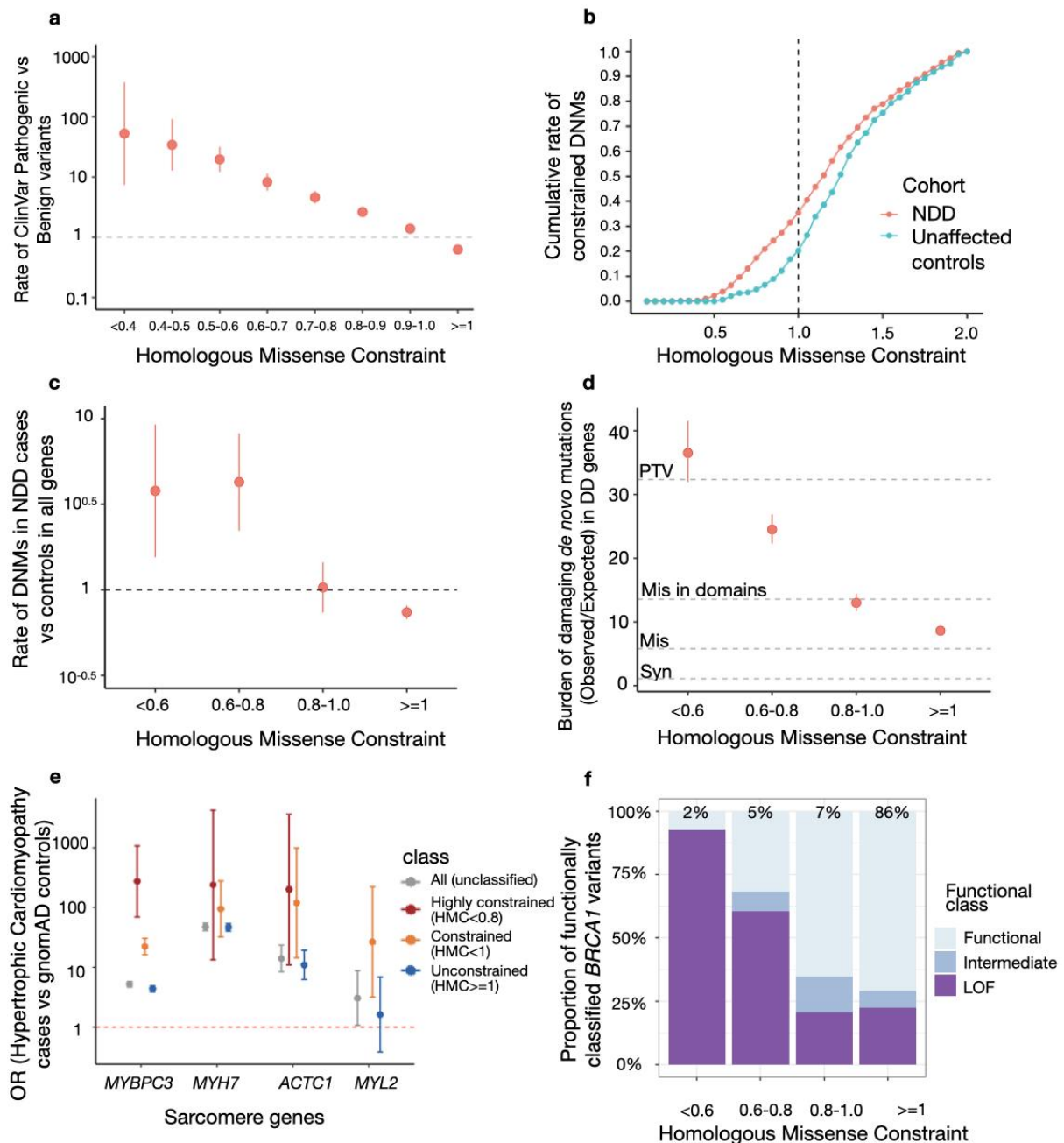
177 We also tested the ability of HMC to predict deleterious variants causing adult-onset
178 disorders. We performed a case-control gene burden test in 6,327 patients with
179 hypertrophic cardiomyopathy from the SHaRe registry²³ using the 125,748 gnomAD
180 v2.1.1 exomes as controls. For four sarcomere genes carrying HMC assessable
181 variants, cases have more HMC constrained variants than controls compared to
182 unconstrained or unclassified variants (**Figure 2e**), although due to low variant
183 numbers this is only individually significant for *MYBPC3* (P -value= 1×10^{-121}). Genes
184 and variants involved in late-onset phenotypes with smaller effects on reproductive
185 fitness often do not show high signals of genetic constraint. These results suggest
186 that selection signals from early-onset disease genes containing homologous
187 domains could also inform critical variants in late-onset disease genes. Applied to
188 individual genes, we expect HMC would have better statistical power in genes with
189 domains from a large Pfam family with more assessable positions as exemplified
190 here: the I-set (Pfam ID: PF07679) and FN3 (Pfam ID: PF00041) domains in

191 *MYBPC3* belong to large domain families with 785 and 597 homologous copies
192 respectively in the exome while domains from the other three tested genes belong to
193 domain families with no more than 72 copies in the exome (the domain from the
194 largest family: EF-hand_1 (Pfam ID: PF00036) in *MYL2*).

195

196 As a further independent evaluation, we compared HMC with functional data from a
197 multiplexed assay of variant effects (MAVE) for *BRCA1*²¹. While HMC shows only
198 modest correlation with *BRCA1* MAVE scores in the continuum space (Pearson
199 correlation coefficient $r = 0.2$, reflecting the limited sensitivity of HMC), 70% of highly
200 constrained (HMC<0.8) *BRCA1* variants are loss-of-function, compared to only 23%
201 of unconstrained positions (association between constraint and *in vitro* functional
202 classification OR=8.9, 95%CI=5.4-14.5; **Figure 2f**).

203



204

205 **Figure 2. HMC accurately distinguishes pathogenic variants from benign**

206 **variants.** (a) Highly constrained positions within protein domains are enriched for

207 pathogenic variants and unconstrained variants are depleted for pathogenic variants.

208 The rate of ClinVar Pathogenic variants vs Benign variants (Risk ratios) within

209 various HMC constrained/unconstrained bins was plotted with 95% CI. Rate of

210 ClinVar Pathogenic vs Benign variants (Rate_{Pathogenic vs Benign}) was calculated as $N_{in\ the}$

211 $bin, pathogenic / N_{total, pathogenic}$ vs $N_{in\ the\ bin, benign} / N_{total, benign}$. A total of 13,009 ClinVar

212 Pathogenic and 3,914 ClinVar benign variants (n=3,914) were assessed. (b)
213 Missense *de novo* mutations observed in a cohort of individuals with
214 neurodevelopmental disorders (NDD; n=5,264) are found at more highly-constrained
215 residues than *de novo* mutations observed in unaffected controls (n=2,179). The
216 cumulative rate of constrained *de novo* mutations in cases ($N_{\text{DNMs with HMC}<X \text{ in cases}}/N_{\text{total DNMs in cases}}$) was plotted to compare with that in controls ($N_{\text{DNMs with HMC}<X \text{ in controls}}/N_{\text{total DNMs in controls}}$). In total, 1,209 DNMs in cases and 337 in controls are
219 assessed in all genes. (c) Missense *de novo* mutations affecting highly constrained
220 domain positions are significantly enriched in NDD cases versus unaffected controls.
221 The rate of DNMs in cases was compared with that in controls in various HMC
222 constrained/unconstrained bins. Rate of DNMs in cases vs controls was calculated
223 as $N_{\text{in the bin, cases}}/N_{\text{total, cases}}$ vs $N_{\text{in the bin, controls}}/N_{\text{total, controls}}$. (d) In 285 genes associated
224 with developmental disorders, HMC prioritises damaging *de novo* missense
225 mutations with a comparable effect size as protein-truncating variants (PTV) in
226 31,058 parent-proband trios of developmental disorders (DD). We compared the
227 prevalence of missense *de novo* mutations (DNM) in established DD-associated
228 genes in individuals with DD against that of expected *de novo* mutations predicted by
229 context-based mutability, and plot the ratio ("burden") for missense DNMs stratified
230 by HMC score. The burden (Obs/Exp) ratio was calculated as $N_{\text{observed DNMs in the bin, in cases}}/N_{\text{expected DNMs in the bin, in cases}}$. As baseline references, the dotted lines show the
232 burden (Obs/Exp) ratio for synonymous DNMs ($OR_{\text{Syn}}=1.1$, 95%CI=1.0-1.2),
233 missense DNMs (without HMC stratification; $OR_{\text{Mis}}=5.8$, 95%CI=5.6-6.0), missense
234 DNMs within annotated Pfam domains ($OR_{\text{Mis in domains}}=13.5$, 95%CI=12.9-14.3), and
235 protein-truncating DNMs ($OR_{\text{PTV}}=32.4$, 95%CI=30.8-34.0). Missense DNMs at the
236 most highly-constrained residues (HMC<0.6) show an association signal similar to

237 that of protein-truncating DNMs. (e) Highly constrained (HMC<0.8) or nominally
238 constrained missense variants (HMC<1) have increased association with
239 hypertrophic cardiomyopathy compared with controls. We calculated the odds of
240 carrying a rare missense variant for individuals with hypertrophic cardiomyopathy,
241 and for the gnomAD reference population, and show the odds ratio for all rare
242 missense variants, and for rare missense variants stratified by HMC scores.
243 Constrained variants in *MYBPC3* are more strongly disease associated. Data are
244 sparser for the other three genes shown, which are much rarer causes of
245 hypertrophic cardiomyopathy, but the trend is concordant. (f) Highly constrained
246 missense variants are more likely to alter protein function in *BRCA1*. Given *BRCA1*
247 variants classified with *in vitro* assays (LOF: loss-of-function; Functional: functionally
248 neutral; Intermediate: intermediate between the two), we show the proportion of each
249 functional class in HMC constrained/unconstrained bins. Constrained bins include
250 more loss-of-function variants compared with unconstrained bins.

251

252

253

254

255

256

257

258

259

260

261

262 We next compared the performance of HMC against existing pathogenicity scores,
263 using ClinVar variants as a reference set. We first compared HMC to existing sub-
264 genic constraint models: Constrained Coding Region⁴ (CCR), Regional Missense
265 Constraint⁵ (RMC), and a homologous-residue-based conservation metric
266 para_zscore²². Using the authors' recommended thresholds, HMC has better
267 precision than all other methods (**Figure 3a**). We also compared HMC to the state-
268 of-the-art supervised meta-predictors: M-CAP²³, MPC⁵, REVEL²⁴ and CADD²⁵. HMC
269 performs comparably to these tools using the highly constrained threshold (<0.8;
270 **Figure 3b**), even though the meta-predictors leverage multiple lines of evidence, and
271 some have been trained directly or indirectly using ClinVar which will tend to inflate
272 their performance, while HMC uses a single line of evidence and is naive to this
273 data.

274

275 We also compared HMC with the above existing pathogenicity scores and constraint
276 models for prioritising deleterious DNMs. Using the 31K DD cohort, HMC
277 outperforms all seven existing tools, being able to identify a subset of DNMs with the
278 highest enrichment in the 285 dominant DD genes (top 5%, **Figures 3c and 3d**),
279 with an effect size as strong as protein-truncating variants. This highlights that HMC
280 is highly effective to distinguish pathogenic variants from bystanders within disease
281 genes.

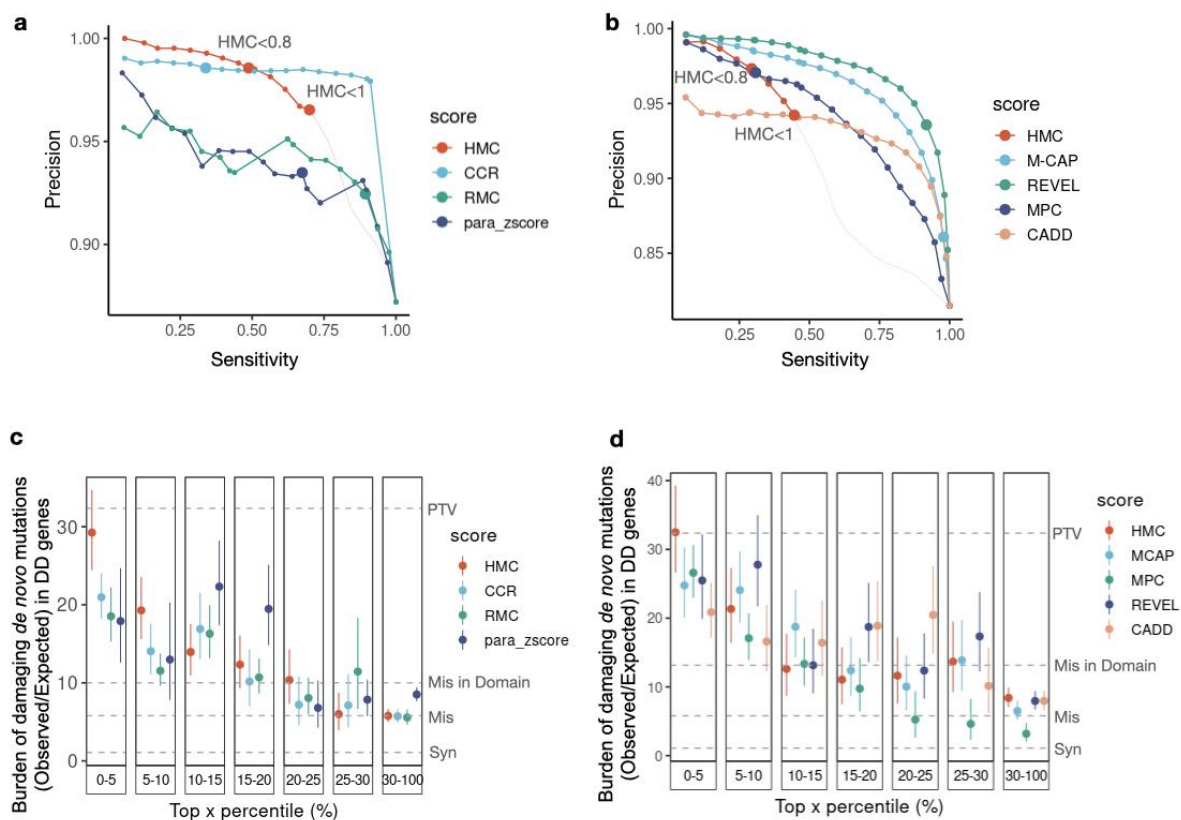
282

283

284

285

286



287

288 **Figure 3. HMC has greater precision than other constraint metrics, and**
 289 **comparable performance to meta-predictor pathogenicity scores. (a)** Using
 290 ClinVar variants, the Precision-Sensitivity curve demonstrates that HMC has higher
 291 precision over the constraint- and homologous-residue-based methods in top-ranked
 292 variants and within authors' recommended thresholds (dots with larger size). The
 293 recommended threshold and the corresponding precision and sensitivity for each
 294 tool is: HMC<0.8 (98.6%, 48.8%), CCR>95 (98.6%, 34%), RMC<0.6(93.5%, 67.4%)
 295 and para_zscore>0 (92.4%, 89.3%). The portion of the curve corresponding to
 296 HMC<1 is shown in red, while HMC >1 is shown in pale pink for completeness. As
 297 each approach generates predictions on different areas of the exome, we analysed
 298 the intersection of ClinVar variants that can be scored by all four methods (3,661
 299 pathogenic and 537 benign variants). (b) HMC has comparable precision as existing
 300 state-of-the-art supervised meta-predictors. Dots with larger size indicate the

301 performances (precision, sensitivity) using authors' recommended thresholds:
302 HMC<0.8 (96.9%, 28.2%), M-CAP>=0.025 (83.6%, 97.2%), REVEL>0.5 (92.2%,
303 90.8%), CADD >=10(No variants are scored as deleterious), MPC>=2 (97.8%,
304 28.6%). The performances are reported using 12,428 ClinVar pathogenic variants
305 and 2,824 benign variants, which could be assessed by all the benchmarked
306 methods. (c-d) Using the 31K trio data of DD, the burden of *de novo* mutations
307 (Observed/Expected) in DD-associated genes is compared to evaluate the precision
308 of predicting damaging variants (the higher the burden, the more likely the variants
309 are damaging). For a given tool, variants are grouped into bins based on their
310 percentile of predicted pathogenicity among all assessed variants. (c) Comparison
311 between HMC with existing constraint-based and homologous residue-based scores
312 using DNM burden; (d) Comparison between HMC with existing state-of-the-art meta
313 learners using DNM burden.

314

315

316

317

318

319

320

321

322

323

324

325

326 After confirming HMC is highly precise compared with existing approaches to
327 evaluating missense constraints, we next ask whether it is orthogonal to them. We
328 assessed the distribution of HMC constrained variants across the existing missense
329 constraint metrics including gene-level constraint (MOEUF)², and sub-genic regional
330 constraint scores (CCR and RMC). Constrained homologous residues detected by
331 HMC are distributed across full ranges of these existing metrics in either constrained
332 or unconstrained genes/regions. If a gene/region is more constrained as a whole, on
333 average it also has more constrained residues compared to a less constrained
334 gene/region (**Figure 4**). However, there are substantial numbers of highly
335 constrained missense variants uniquely classified by HMC (<0.8): 893,063 not
336 prioritised by either of the sub-genic metrics, of which 351,175 are not prioritised by
337 any of the existing metrics.

338

339

340

341

342

343

344

345

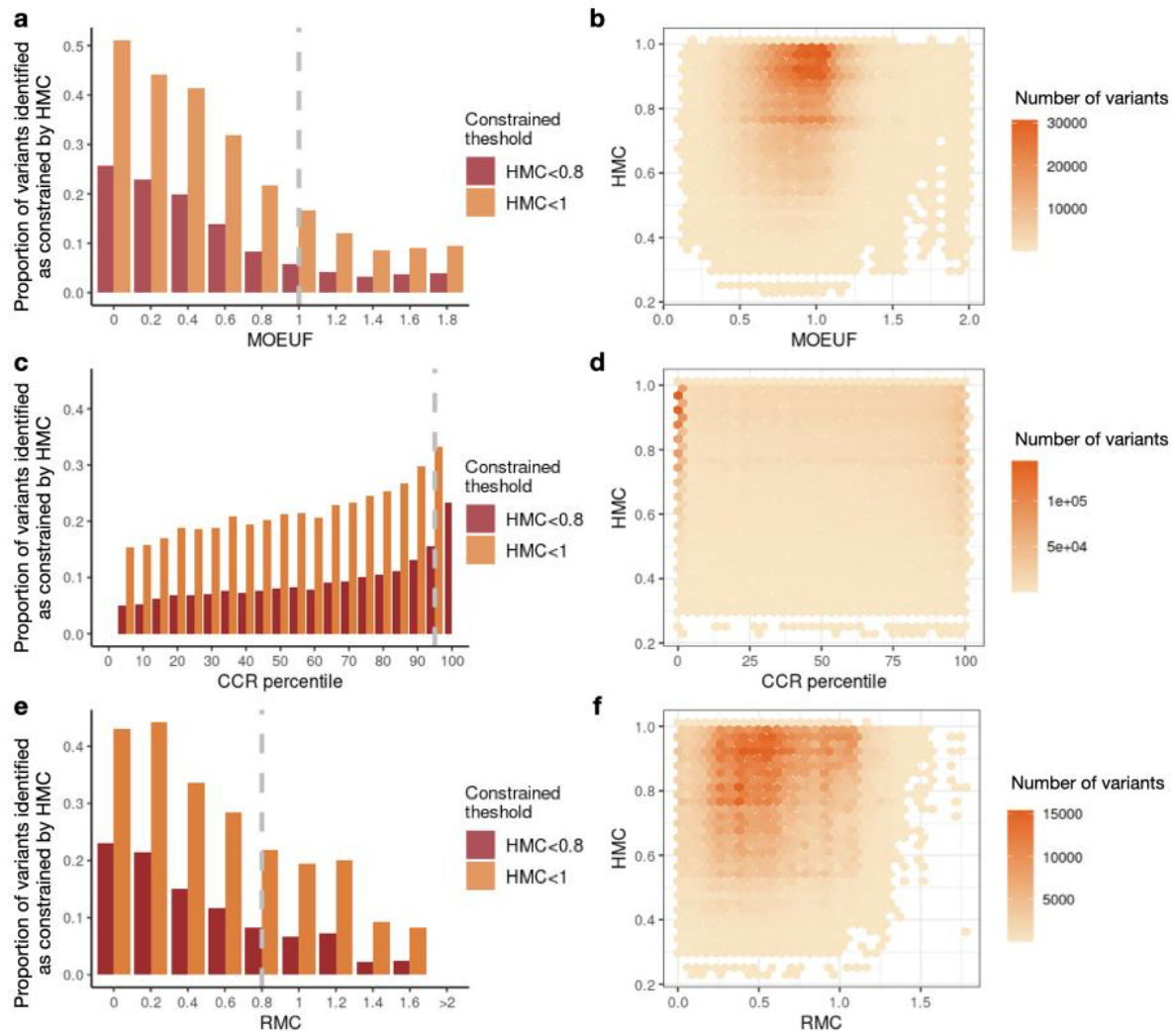
346

347

348

349

350



351

352

Figure 4. Comparing the distributions of HMC score with existing gene-level

353

and regional level constraint scores. Here we show that HMC is not co-linear with

354

other metrics, and therefore is likely to provide additional information when used in

355

combination. Bar plots in the first column display the proportion of variants identified

356

as constrained by HMC across genes or regions ($N_{\text{HMC constrained variants in the bin}}/N_{\text{variants in the bin}}$).

357

Providing a further detailed view of the bar plots, 2D-histogram plots in the

358

second column display the number of HMC constrained variants within various

359

ranges across gene/region constraint scores. (a-b) The relationship between HMC

360

and a gene's MOEUF score (gene-level constraint of missense variants; a lower

361

value indicates higher constraint). A gene with $\text{MOEUF} \geq 1$ (grey dashed line) is

362 considered as nominally unconstrained. (c-d) The relationship between HMC and
363 CCR (a higher percentile indicates higher constraint). A region with CCR percentile
364 <95% (grey dashed line) is considered as unconstrained recommended by authors⁴.
365 (e-f) The relationship between HMC and RMC (a lower value indicates higher
366 constraint). A region with RMC>0.8 (grey dashed line) is considered unconstrained
367 as recommended by authors.

368

369

370

371

372

373

374

375

376

377

378

379

380

381

382

383

384

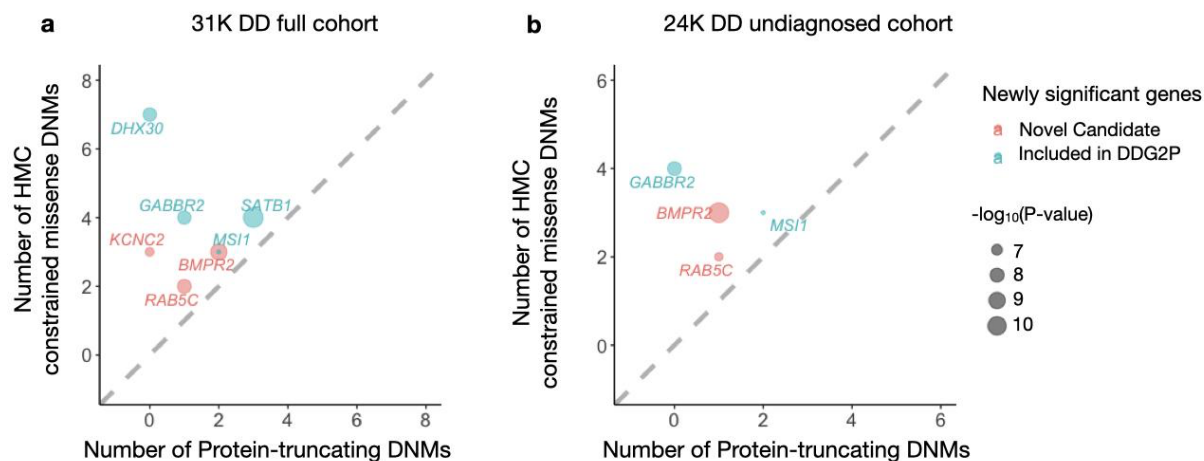
385

386

387 We further investigate whether HMC could improve gene discovery in developmental
388 disorders given that HMC represents an orthogonal measure of variant
389 deleteriousness and is highly precise to predict disease-causing DNMs in DD genes.
390 HMC prioritises a subset of missense DNMs that show a significant excess burden in
391 the 31K DD probands, in genes that are not previously known to be associated with
392 DD (i.e. not currently considered as diagnostic; HMC<0.8: Obs/Exp=1.37
393 95%CI=1.25-1.51; **Figure S5**), suggesting its potential to discover unknown DD
394 genes. We updated a gene-specific *de novo* enrichment test (DeNovoWEST)¹⁸ by
395 incorporating HMC to weight missense variants (see **Methods**) in the 31K DD
396 cohort. Consequently, we observed increased DNM burden of upweighted variants
397 and decreased DNM burden of downweighted ones, indicating the improved
398 separation of pathogenic from benign variants (**Table S1**). Our upgraded tests
399 identified 286 disease-associated genes in the full cohort, and 97 in those previously
400 undiagnosed (probands who do not carry pathogenic variants in consensus
401 diagnostic genes, as previously defined¹⁸) at genome-wide significance (Bonferroni
402 adjusted P -value<0.05/(2×18,762)).
403
404 Compared with the original study¹⁸, there are seven newly significant genes across
405 the two tests, which carry at least one constrained missense variant, confirming that
406 their elevated significance signal is driven by HMC (**Table S2-S3**). Four of these
407 genes have previously been published in association with DD via other lines of
408 evidence and are currently included in the Developmental Disorders Genotype-to-
409 Phenotype Database²⁶(DDG2P), indicating that our results provide independent
410 support about their gene-disease association. Three of these genes (*BMP2*,
411 *KCNC2* and *RAB5C*) have not yet been included in the DDG2P. *BMP2* is known to

412 cause pulmonary arterial hypertension^{27,28}. *KCNC2* has been independently
413 suggested to be a new candidate epilepsy gene^{29–31}. Importantly, the newly
414 significant genes all have more constrained missense DNMs than protein-truncating
415 DNMs, suggesting the potential involvement of a gene-function altering mechanism
416 **(Figure 5)**.

417
418 Here we have described a novel measure of genetic intolerance, HMC, to predict
419 deleterious missense variants. Compared with existing metrics that aggregate
420 variants over “horizontal” regions of the genome, HMC considers “vertical” space
421 across homologous regions, enabling us to assess genetic constraint at the single
422 amino-acid level. We demonstrate how HMC is highly precise and orthogonal to
423 existing computational evidence, thus can be broadly used to interpret missense
424 variants and to enhance gene discovery. Though the signal is expected to be driven
425 by genes associated with early-onset disorders under strong negative selection,
426 HMC is informative in adult-onset disease genes, likely leveraging constraint signals
427 from early-onset genes with homologous domains. Although HMC can currently only
428 assess ~20% of the exome, its statistical power will be increased with the ongoing
429 growth of large-scale population genomics data, efforts of protein family
430 classification, and development of computational structural genomics (providing an
431 alternative definition of homologous variants). HMC provides a powerful new tool for
432 the interpretation of genetic variation.



433

434 **Figure 5. *De novo* variants identified in 31,058 parent-proband trios reveal**
435 **seven genes associated with developmental disorders at genome-wide**
436 **significance for the first time in the full DD cohort (a) and the previously-**
437 **undiagnosed subset (b).** Four of these genes have been previously curated as DD
438 genes on the basis of other lines of evidence, and are already included in the G2P
439 database as established Developmental Disorder genes (blue), while three genes
440 represent new candidate DD genes (red). Numbers of constrained missense DNMs
441 classified by HMC and protein-truncating DNMs were compared. The newly-
442 significant associated genes likely act through altered function mechanisms as there
443 are more constrained missense variants than PTVs.

444

445

446

447

448

449

450

451

452 **Code availability**

453 The essential scripts used to generate HMC scores and recreate the figures in the
454 main text are available at [https://github.com/ImperialCardioGenetics/homologous-](https://github.com/ImperialCardioGenetics/homologous-missense-constraint)
455 [missense-constraint](https://github.com/ImperialCardioGenetics/homologous-missense-constraint).

456

457 **Data availability**

458 HMC scores for all assessable variants are available via www.cardiodb.org/hmc.

459 External data used in the study were obtained from the following approaches/URLs:

460 The IDs of RefSeq Select Transcripts were downloaded from the UCSC Genome

461 Browser using the Table Browser tool (downloaded date: Nov 28th 2020; options

462 used: group - “Genes and Gene Predictions”, track - “NCBI RefSeq”, table - “RefSeq

463 Select”); ClinVar,

464 https://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh37/archive_2.0/2020/clinvar_20201

465 [114.vcf.gz](https://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh37/archive_2.0/2020/clinvar_20201114.vcf.gz) (downloaded at Nov 2020) ; Developmental Disorder Genotype-

466 Phenotype Database (DDG2P), <https://www.deciphergenomics.org/ddd/ddgenes>

467 (version 2021.11.05); The 31K DD trio data and the original DeNovoWEST,

468 <https://github.com/HurlesGroupSanger/DeNovoWEST>; gnomAD exome, v2.1.1,

469 <https://gnomad.broadinstitute.org/downloads>; The hypertrophic cardiomyopathy case

470 series curated by the SHaRe Consortium (data release 2019Q3):

471 https://github.com/ImperialCardioGenetics/CardioBoost_manuscript/tree/master/data

472 [/cardiomyopathy/share_variant_count.RData](https://github.com/ImperialCardioGenetics/CardioBoost_manuscript/tree/master/data/cardiomyopathy/share_variant_count.RData); CCR score, [https://github.com/quinlan-](https://github.com/quinlan-lab/ccr)

473 [lab/ccr](https://github.com/quinlan-lab/ccr); RMC and MPC score, [https://storage.googleapis.com/gcp-public-data--](https://storage.googleapis.com/gcp-public-data--gnomad/legacy/exac_browser/regional_missense_constraint.tsv)

474 [gnomad/legacy/exac_browser/regional_missense_constraint.tsv](https://storage.googleapis.com/gcp-public-data--gnomad/legacy/exac_browser/regional_missense_constraint.tsv),

475 ftp://ftp.broadinstitute.org/pub/ExAC_release/release1/regional_missense_constraint/

476 [fordist_constraint_official_mpc_values.txt.gz](ftp://ftp.broadinstitute.org/pub/ExAC_release/release1/regional_missense_constraint/fordist_constraint_official_mpc_values.txt.gz); M-CAP score,

477 <http://bejerano.stanford.edu/mcap/> (v1.4); REVEL score,
478 <https://sites.google.com/site/revelgenomics>; CADD,
479 <https://cadd.gs.washington.edu/download>; para_zscore,
480 https://zenodo.org/record/3582386#.YYxNXb1_rSw (version 9).

481

482 **Acknowledgements**

483 We thank Matt Hurles for helpful background discussions. This work was supported
484 by Medical Research Council (UK), British Heart Foundation [RE/18/4/34215], the
485 NIHR Imperial College Biomedical Research Centre, and the Wellcome Trust
486 [107469/Z/15/Z, 200990/A/16/Z]. NW is currently supported by a Sir Henry Dale
487 Fellowship jointly funded by the Wellcome Trust and the Royal Society (Grant Number
488 220134/Z/20/Z) and funding from the Rosetrees Trust.

489

490 For the purpose of open access, the author has applied a CC BY public copyright
491 licence to any Author Accepted Manuscript version arising from this submission. The
492 views expressed in this work are those of the authors and not necessarily those of the
493 funders.

494

495 **References**

- 496 1. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**,
497 285–291 (2016).
- 498 2. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in
499 141,456 humans. *Nature* **581**, 434–443 (2020).
- 500 3. Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S. & Goldstein, D. B. Genic intolerance
501 to functional variation and the interpretation of personal genomes. *PLoS Genet.* **9**,

- 502 e1003709 (2013).
- 503 4. Havrilla, J. M., Pedersen, B. S., Layer, R. M. & Quinlan, A. R. A map of constrained
504 coding regions in the human genome. *Nat. Genet.* **51**, 88–95 (2019).
- 505 5. Samocha, K. E. *et al.* Regional missense constraint improves variant deleteriousness
506 prediction. doi:10.1101/148353.
- 507 6. Whiffin, N. *et al.* Characterising the loss-of-function impact of 5' untranslated region
508 variants in 15,708 individuals. *Nature Communications* vol. 11 (2020).
- 509 7. Short, P. J. *et al.* De novo mutations in regulatory elements in neurodevelopmental
510 disorders. *Nature* **555**, 611–616 (2018).
- 511 8. Vitsios, D., Dhindsa, R. S., Middleton, L., Gussow, A. B. & Petrovski, S. Prioritizing non-
512 coding regions based on human genomic constraint and sequence context with deep
513 learning. *Nat. Commun.* **12**, 1504 (2021).
- 514 9. Strumillo, M. J. *et al.* Conserved phosphorylation hotspots in eukaryotic protein domain
515 families. *Nat. Commun.* **10**, 1977 (2019).
- 516 10. Mistry, J., Bateman, A. & Finn, R. D. Predicting active site residue annotations in the
517 Pfam database. *BMC Bioinformatics* **8**, 298 (2007).
- 518 11. Wiel, L., Venselaar, H., Veltman, J. A., Vriend, G. & Gilissen, C. Aggregation of
519 population-based genetic variation over protein domain homologues and its potential
520 use in genetic diagnostics. *Hum. Mutat.* **38**, 1454–1463 (2017).
- 521 12. Ware, J. S., Walsh, R., Cunningham, F., Birney, E. & Cook, S. A. Paralogous annotation
522 of disease-causing variants in long QT syndrome genes. *Hum. Mutat.* **33**, 1188–1191
523 (2012).
- 524 13. Walsh, R., Peters, N. S., Cook, S. A. & Ware, J. S. Parologue annotation identifies novel
525 pathogenic variants in patients with Brugada syndrome and catecholaminergic
526 polymorphic ventricular tachycardia. *J. Med. Genet.* **51**, 35–44 (2014).
- 527 14. Wiel, L. *et al.* MetaDome: Pathogenicity analysis of genetic variants through aggregation
528 of homologous human protein domains. *Hum. Mutat.* **40**, 1030–1038 (2019).
- 529 15. El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**,

- 530 D427–D432 (2019).
- 531 16. Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology
532 Information. *Nucleic Acids Res.* **49**, D10–D17 (2021).
- 533 17. Satterstrom, F. K. *et al.* Large-Scale Exome Sequencing Study Implicates Both
534 Developmental and Functional Changes in the Neurobiology of Autism. *Cell* **180**, 568–
535 584.e23 (2020).
- 536 18. Kaplanis, J. *et al.* Evidence for 28 genetic disorders discovered by combining healthcare
537 and research data. *Nature* **586**, 757–762 (2020).
- 538 19. Ware, J. S., Samocha, K. E., Homsy, J. & Daly, M. J. Interpreting de novo Variation in
539 Human Disease Using denovolyzeR. *Curr. Protoc. Hum. Genet.* **87**, 7.25.1–7.25.15
540 (2015).
- 541 20. Samocha, K. E. *et al.* A framework for the interpretation of de novo mutation in human
542 disease. *Nat. Genet.* **46**, 944–950 (2014).
- 543 21. Findlay, G. M. *et al.* Accurate classification of BRCA1 variants with saturation genome
544 editing. *Nature* **562**, 217–222 (2018).
- 545 22. Lal, D. *et al.* Gene family information facilitates variant interpretation and identification of
546 disease-associated genes in neurodevelopmental disorders. *Genome Med.* **12**, 28
547 (2020).
- 548 23. Jagadeesh, K. A. *et al.* M-CAP eliminates a majority of variants of uncertain significance
549 in clinical exomes at high sensitivity. *Nat. Genet.* **48**, 1581–1586 (2016).
- 550 24. Ioannidis, N. M. *et al.* REVEL: An Ensemble Method for Predicting the Pathogenicity of
551 Rare Missense Variants. *Am. J. Hum. Genet.* **99**, 877–885 (2016).
- 552 25. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human
553 genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
- 554 26. Thormann, A. *et al.* Flexible and scalable diagnostic filtering of genomic variants using
555 G2P with Ensembl VEP. *Nat. Commun.* **10**, 2373 (2019).
- 556 27. Lane, K. B. *et al.* Heterozygous germline mutations in BMPR2, encoding a TGF- β
557 receptor, cause familial primary pulmonary hypertension. *Nature Genetics* vol. 26 81–84

558 (2000).

559 28. Rehm, H. L. *et al.* ClinGen — The Clinical Genome Resource. *New England Journal of*
560 *Medicine* vol. 372 2235–2242 (2015).

561 29. Vetri, L. *et al.* A de novo heterozygous mutation in KCNC2 gene implicated in severe
562 developmental and epileptic encephalopathy. *Eur. J. Med. Genet.* **63**, 103848 (2020).

563 30. Ryzdanicz, M. *et al.* A recurrent de novo variant supports KCNC2 involvement in the
564 pathogenesis of developmental and epileptic encephalopathy. *Am. J. Med. Genet. A*
565 **185**, 3384–3389 (2021).

566 31. Rademacher, A. *et al.* Whole-Exome Sequencing in NF1-Related West Syndrome
567 Leads to the Identification of KCNC2 as a Novel Candidate Gene for Epilepsy.
568 *Neuropediatrics* **51**, 368–372 (2020).

569

570

571

572

573

574

575

576

577

578

579

580 **Methods**

581 **Identification of homologous residues from domain family alignments**

582 The family alignments of all 6,196 human protein domains generated using the NCBI
583 RefSeq sequence database were downloaded from Pfam¹⁵ (Pfam-A.full.ncbi.gz
584 version 32.0). Given a multiple sequence alignment of a domain family, amino acids
585 in the same column of the alignment were considered homologous.

586

587 **Annotation of molecular consequences of variants**

588 RefSeq Select transcripts were used throughout the whole analysis such that each
589 protein-coding gene has a single high-quality representative transcript. The
590 consequences of variants were annotated by VEP (release 101)³². Only single-
591 nucleotide variants with VEP annotated as “missense_variant” were included in the
592 analysis.

593

594 **Developing a selection-neutral, sequence-context mutational model**

595 To estimate the number of substitutions expected on a single nucleotide, we
596 constructed a neutral mutational model using the gnomAD reference population.
597 Previous studies have shown that the mutation rate of single nucleotide substitution
598 under neutral selection could be predicted based on sequence context and
599 methylation level¹. Given the baseline substitution rate using a tri-nucleotide
600 sequence text model estimated from variants in intergenic or intronic regions by
601 gnomAD², we calibrated the baseline mutation rate to probabilities of synonymous
602 variants (presumed to be neutral substitutions) within the 125,478 exomes in
603 gnomAD following the procedures described in the gnomAD flagship paper².

604

605 We firstly used linear regression to predict the proportions of neutral substitutions
606 given the baseline mutation rates. For each possible tri-nucleotide sequence context,
607 the proportion of neutral substitutions is calculated as the ratio of observed
608 synonymous substitutions over all possible synonymous substitutions. For example,
609 to calculate the proportion of neutral substitutions from AAT to AGT, we firstly find
610 the number of all possible synonymous variants introduced by mutating AAT to AGT
611 along the exome and then count those observed in gnomAD v2 exome data. This
612 ratio of “observed” to “all possible” is used as the dependent variable in linear
613 regression. Since the observation of substitutions would be biased by sequencing
614 coverage, at this step only sites with high coverage (median depth 40) are included
615 in the regression. Two linear regression models were fitted, one for substitutions at
616 CpG sites and the other for non-CpG sites (**Figure S2**). The methylation data for
617 CpG sites was downloaded from gnomAD public datasets and was categorised into
618 three bins: low, medium, and high methylation levels as previously described². With
619 these predicted probabilities of substitutions, we can estimate the expected number
620 of single-nucleotide variants under neutral selection (Expected) in the 125,478
621 exomes in gnomAD.

622

623 Secondly, we adjusted the probabilities of neutral substitutions for low-coverage sites
624 (median depth<40). To this end, the Observed/Expected ratios for synonymous
625 variants were aggregated for each sequencing coverage (measured by median
626 coverage among gnomAD samples). Given a sequencing coverage, it was
627 calculated as: the expected number of variants is the sum of predicted proportions of
628 neutral substitutions for each site derived from the first step, indicating the number
629 we expect with high-coverage sequencing; the observed number of variants is the

630 sum of observed synonymous variants for each site. A linear model was fitted to
631 predict the Observed/Expected ratios given a sequencing coverage on a \log_{10} scale
632 ($R^2=0.96$, $P\text{-value}<2.2\times 10^{-16}$; **Figure S3**). The predicted Observed/Expected ratios
633 by the model were used as correction factors to adjust the expected number of
634 variants at low-coverage sites.

635

636 **Estimating Homologous Residue Constraint**

637 An overview of measuring Homologous Residue Constraint is illustrated in **Figure 1**.
638 For an aligned position in a Pfam domain family, we assessed all possible missense
639 substitutions. Among all the possible missense substitutions, the number of
640 substitutions directly observed in gnomAD was counted (Observed). The expected
641 number of missense substitutions was calculated as the sum of predicted
642 probabilities of substitutions given by the neutral mutational model (Expected). The
643 genetic intolerance of this aligned position was calculated as the ratio of
644 Observed/Expected.

645

646 Given the variability in the number of aligned domains, in order to control the quality
647 of assessing genetic constraint in homologous residues, we excluded any domain
648 position with less than three expected variants as the number of possible missense
649 variants that occurred at this residue position is too small for us to evaluate genetic
650 constraint robustly. If the number of observed substitutions follows a Poisson
651 distribution under the null hypothesis (no selection), even with zero observed
652 substitutions, the expected number needs to be at least three to reach the
653 significance threshold (the probability of observing zero occurrences with mean
654 occurrence as three is 0.049. In R, it is calculated as “`ppois(0,3)=0.049`”). It might

655 also indicate the corresponding column is constructed with low confidence filled with
656 a large proportion of gaps (>95% in our observation). Filtering these columns might
657 also limit the effect of alignment bias on defining homologous residues.

658

659 Homologous Residue Constraint is defined as the upper limit of 95% confidence
660 interval for the Observed/Expected ratio. The confidence interval for the
661 Observed/Expected ratio was estimated using a Bayesian approach as previously
662 described². The unknown true Observed/Expected ratio (constraint) was considered
663 as a random variable with a uniform prior between 0 and 2 (in computing, discretized
664 into a sequence of 2000 values from 0 to 2, incremented by 0.001). The likelihood
665 function for a given constraint value is given as the Poisson density:

666

$$667 \quad Pr(X = Observed | constraint = \lambda) = \frac{(\lambda * Expected)^{Observed} e^{-\lambda * Expected}}{Observed!}.$$

668

669 Thus, the posterior probability of a given constraint value could be derived by:

670

$$671 \quad Pr(constraint = \lambda | Observed, Expected) = \frac{Pr(constraint = \lambda) * Pr(X = Observed | constraint = \lambda)}{\sum_{\lambda} Pr(constraint = \lambda) * Pr(X = Observed | constraint = \lambda)}.$$

672

673 We could further obtain the 95% confidence interval of constraint by taking the 2.5%
674 and 97.5% quantile from its posterior probability distribution. Therefore, the upper
675 bound of 95% CI is taken as the constraint score of homologous residues (HMC). If a
676 residue is scored as HMC < 1, it indicates that missense variants disrupting the given
677 domain position are significantly (P -value < 0.05) depleted of variants thus under
678 selection pressure.

679

680 There are 28,032,394 (~40% of 70 million possible missense variants) rare (gnomAD
681 MAF<0.1%) missense variants in 15,305 genes (out of 19,212 genes) overlapping
682 5,807 Pfam domains. After excluding domain positions with limited statistical power,
683 there are 15,236,101 possible rare missense variants from 699 Pfam families with
684 78,070 domain positions assessable in 9,918 genes. We identified 3,304,332
685 possible missense variants (21.7% of assessable) at constrained (HMC<1) positions
686 in 596 Pfam domains. 1,322,835 possible missense variants (8% of assessable)
687 were identified at highly constrained residues (HMC<0.8) of 458 domains.

688

689 **Evaluating the pathogenicity of ClinVar variants**

690 The association of HMC with known disease-causing variants was tested using
691 ClinVar³³. The VCF file was downloaded from the ClinVar public FTP site (version
692 20201114
693 [https://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh37/archive_2.0/2020/clinvar_20201](https://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh37/archive_2.0/2020/clinvar_20201114.vcf.gz)
694 [114.vcf.gz](https://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh37/archive_2.0/2020/clinvar_20201114.vcf.gz)). We extracted 22,886 Pathogenic/Likely pathogenic missense variants in
695 Pfam domains, whose clinical significance was recorded as “Pathogenic”,
696 “Likely_pathogenic” or “Pathogenic/Likely_pathogenic”. 7,137 Benign/Likely benign
697 variants in Pfam domains were extracted with clinical significance recorded as
698 “Benign”, “Likely_benign” or “Benign/Likely_benign”. After keeping the HMC
699 assessable domain positions, 13,009 Pathogenic/Likely pathogenic and 3,914
700 Benign/Likely benign variants were used as test data. Only variants with no
701 conflicting interpretation were included in the test set.

702

703 **Evaluating the pathogenicity of *de novo* variants**

704 To test the enrichment of *de novo* variants (DNMs) prioritised by HMC in affected
705 individuals versus unaffected individuals, we analysed the published DNMs in 5,264
706 patients ascertained with neurodevelopmental disorders, 6,430 patients ascertained
707 with autism spectrum disorder, and 2,179 unaffected controls curated by Satterstrom
708 *et.al*¹⁷.

709

710 We applied an independent approach to measure the accuracy of predicting
711 damaging missense *de novo* mutations by testing the enrichment of DNMs prioritised
712 by HMC in affected individuals versus neutral variants estimated by a null sequence-
713 context based *de novo* mutational model²⁰. This measurement can also be used to
714 assess whether HMC could distinguish pathogenic and benign variants within
715 disease genes as the enrichment of DNMs in cases vs control individuals can be
716 driven by gene-level disease association. We analysed the published DNMs in
717 31,058 patients with developmental disorders. The burden of DNMs was calculated
718 as the ratio of the number of observed DNMs to the number of expected DNMs. The
719 number of observed DNMs was directly counted from the variants seen in the cohort.
720 The number of expected DNMs under neutral selection for the cohort is calculated by
721 summing the product of the trinucleotide *de novo* mutation rate and the number of
722 exome samples ($2 \times 31,058$) for each nucleotide. The *de novo* mutation rate was
723 downloaded from the GitHub repository of denovolyzeR¹⁹
724 ([https://github.com/jamesware/denovolyzeR-ProbabilityTables/blob/master/data-](https://github.com/jamesware/denovolyzeR-ProbabilityTables/blob/master/data-raw/fordist_1KG_mutation_rate_table.txt)
725 [raw/fordist_1KG_mutation_rate_table.txt](https://github.com/jamesware/denovolyzeR-ProbabilityTables/blob/master/data-raw/fordist_1KG_mutation_rate_table.txt)). The effective sample size for X-
726 chromosome is adjusted considering sex-chromosome transmission as previously
727 described¹⁸. Assuming the number of observed DNMs follows a Poisson distribution,
728 the 95% confidence interval for the mean number of observed DNMs could be

729 estimated by using an exact method. In R, it is calculated as “`poisson.test(n_obs,`
730 `conf.level=0.95)`”.

731

732 To be noticed, the set of DNMs published in Satterstrom *et.al*¹⁷ was a compilation of
733 DNMs from previous publications. Since Satterstrom *et.al*¹⁷ have harmonised the
734 quality control for 5.2K cases and 2.2K unaffected controls thus we used this dataset
735 for DNM case-control analysis. We used the larger set of 31K cases in the DNM
736 enrichment analysis. Of note, the 5.2K cases are a subset of 31K cases^{17,18,34} but we
737 think this won't affect the validity of our results: the controls are different in the two
738 analyses. 5.2K is compared with unaffected individuals while 31K is compared with
739 DNM null model. Therefore, even if we fully reuse the same cases, these two
740 analyses shall be considered as independent validations.

741

742 **Testing improving power of gene discovery**

743 To demonstrate the utility of applying HMC to discover more disease genes reliably,
744 we upgraded the gene-specific *de novo* weighted enrichment simulation test
745 (DeNovoWEST)¹⁸ by adding HMC to score missense variants. In the original
746 framework of DeNovoWEST, the weight of a missense variant used in the simulation
747 test depends on the regional missense constraint (RMC). Here we incorporated
748 HMC into this framework through the following procedures: (1) combining HMC with
749 regional missense constraint to label constrained missense variants, thus a
750 missense variant is considered as constrained if either RMC or HMC ($HMC < 0.8$)
751 score it as constrained. Compared with the original version, there are 732,404 more
752 missense variants classified as constrained; (2) updating the weights of missense
753 variants used in DeNovoWEST: we calculated the burden of *de novo* missense

754 variants against a null *de novo* mutational model²⁰ and inferred the corresponding
755 positive predictive values (PPV) for all possible categories using constraint (based
756 on step 1) and CADD scores. The newly derived PPV is used as weights in the
757 downstream gene-specific test. The upgraded test was applied separately in the full
758 (n=31,058) and undiagnosed (n=24,288) cohort of parent-proband trios of
759 developmental disorders¹⁸. Compared with the original version, the DNM burden of
760 constrained missense variants have increased in the 31K DD cohort while the one of
761 unconstrained missense variants have decreased (**Table S1**). This suggests that
762 HMC has improved the classification of pathogenic and benign missense DNMs in
763 the cohort.

764

765 We define newly-significant associated genes driven by HMC scoring that meet all
766 the following criteria: (1) only reached genome-wide significance threshold
767 (Bonferroni adjusted P -value $<0.05/(2 \times 18,762)$) in our upgraded test; (2) carries at
768 least one highly constrained missense variant prioritized by HMC (HMC <0.8). In our
769 upgraded tests, there are seven genes in total that reached genome-wide
770 significance level but did not in the original DeNovoWEST tests: *BMP2R*, *DHX30*,
771 *GABBR2*, *KCNC2*, *MSI1*, *RAB5C*, *SATB1* (**Table S2-S3**). All of them have at least
772 one HMC constrained missense variant thus are defined as newly-significant
773 associated genes in our analysis. There are three genes (*ARHGEF9*, *SETD1B* and
774 *CNKSR2*) that were significant in the original tests but did not in the upgraded tests.
775 Since the significance levels of these three genes before and after the upgrade are
776 in the same scale, we think it's likely due to random variations in different runs of
777 simulations. For those currently not included in DDG2P, we define them as novel
778 candidate genes associated with DD including *BMP2R*, *KCNC2* and *RAB5C*.

779

780 **Compare HMC with functional assay data**

781 We collected mutagenesis experiment data on missense SNVs of 13 exons of
782 *BRCA1* RING domain²¹. Variants are functionally classified as: functionally-neutral,
783 intermediate, loss-of-function. We were able to analyse 1321 out of 3893 missense
784 SNVs in the assay data.

785

786 **References for Methods**

- 787 32. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
- 788 33. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting
789 evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).
- 790 34. Firth, H. V., Wright, C. F. & DDD Study. The Deciphering Developmental Disorders
791 (DDD) study. *Dev. Med. Child Neurol.* **53**, 702–703 (2011).

792