

1 **Cystic fibrosis risk variants confer protection against inflammatory bowel disease**

2

3 **Authors**

4 Mingrui Yu^{1,2}, Qian Zhang³, Kai Yuan^{1,2}, Aleksejs Sazonovs³, Christine Stevens^{1,2,4}, Laura
5 Fachal³, the International Inflammatory Bowel Disease Genetics Consortium, Carl A.
6 Anderson^{3,*}, Mark J. Daly^{1,2,4,*}, Hailiang Huang^{1,2,*}

7

8 ¹Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA

9 ²Stanley Center for Psychiatric Research, the Broad Institute of MIT and Harvard, Cambridge,
10 MA, USA

11 ³Genomics of Inflammation and Immunity Group, Human Genetics Programme, Wellcome
12 Sanger Institute, Wellcome Genome Campus, Hinxton, UK.

13 ⁴Program in Medical and Population Genetics, The Broad Institute of MIT and Harvard,
14 Cambridge, MA, USA

15

16 *These authors jointly supervised this work. Emails: ca3@sanger.ac.uk,
17 mjdaly@broadinstitute.org, hhuang@broadinstitute.org

18

19 **Abstract**

20 Genetic mutations that yield defective cystic fibrosis transmembrane regulator (*CFTR*) protein
21 cause cystic fibrosis, a life-limiting autosomal recessive Mendelian disorder. A protective role of
22 *CFTR* loss-of-function mutations in inflammatory bowel disease (IBD) has been suggested, but
23 its evidence has been inconclusive and contradictory. Here, leveraging the largest IBD exome
24 sequencing dataset to date, comprising 38,558 cases and 66,945 controls in the discovery stage,
25 and 35,797 cases and 179,942 controls in the replication stage, we established a protective role of
26 CF-risk variants against IBD based on evidence from the association test of *CFTR* delF508 (p-
27 value=8.96E-11) and the gene-based burden test of CF-risk variants (p-value=3.9E-07).
28 Furthermore, we assessed variant prioritization methods, including AlphaMissense, using
29 clinically annotated CF-risk variants as the gold standard. Our findings highlight the critical and
30 unmet need for effective variant prioritization in gene-based burden tests.

31

32 Genetic mutations that yield defective cystic fibrosis transmembrane regulator (CFTR) protein
33 are known to cause cystic fibrosis (CF)¹, one of the most common life-threatening autosomal
34 recessive genetic disorders among individuals of European ancestry. *CFTR* is a transmembrane
35 chloride ion channel gene highly expressed in epithelial cells². Defective *CFTR* protein located
36 in the epithelial cell membrane results in defective chloride ion transport³, which leads to water
37 depletion at the cell surface and thick mucus buildup in organs such as the lung, pancreas, sweat
38 gland, and gut. Cellular signaling pathways through which the innate immune system elicits
39 proper immune responses are dysregulated in CF⁴⁻¹⁰, leading to hyperinflammation. However, it
40 is unclear whether *CFTR* mutations are intrinsically pro-inflammatory or whether the heightened
41 inflammation observed in CF results from a heightened response to exacerbated pathogen
42 exposure¹¹⁻¹⁴.

43
44 Inflammatory bowel disease (IBD) is a disorder with chronic inflammation in the gastrointestinal
45 tract. Although earlier research suggested that loss-of-function variants in the *CFTR* gene may
46 lead to a lower prevalence of intestinal inflammation¹⁵ and potentially protect against IBD,
47 evidence has been inconsistent with studies showing conflicting results - some linking these
48 variants to increased risk, others suggesting protection, and some finding no effect at all¹⁶⁻¹⁹.
49 Here, leveraging the largest IBD exome sequencing dataset to date, comprising 38,558 cases and
50 66,945 controls in the discovery stage, and 35,797 cases and 179,942 controls in the replication
51 stage, we provide conclusive evidence that CF-risk variants in *CFTR* confer a protective effect
52 against IBD, and demonstrated the importance of variant prioritization in gene-based burden
53 tests.

54 55 **RESULTS**

56 **Study subjects.** *Broad Institute (Discovery)*: Exome sequencing was performed at the Broad
57 Institute. Study subjects were recruited from different centers and shared with the International
58 Inflammatory Bowel Disease Genetics Consortium (IIBDGC) for 38,558 IBD cases and 66,945
59 controls after quality control analysis (Table 1). Among the cases, 21,478 have Crohn's disease
60 (CD), 14,353 have ulcerative colitis (UC), and 2,727 have IBD unclassified (IBD-U). *Sanger*
61 *Institute (Replication)*: Additional exome and whole genome sequencing were undertaken at the
62 Sanger Institute. Whole exome sequencing was performed on 10,722 CD, 13,147 UC, and 6,211
63 IBD-U, which were matched with whole exome sequencing data obtained from 168,100 controls
64 from the UK Biobank. Extensive quality control was undertaken to harmonize the case and
65 control exome sequencing data (Methods). Whole genome sequencing was performed on another
66 independent sample of 5,717 CD and 11,842 controls. Details on sequencing and data quality-
67 control protocols are described in Methods.

68

Center	Stage	Method	Controls	CD	UC	IBD-U
Broad	Discovery	WES	66,945	21,478	14,353	2,727

Sanger	Replication	WES	168,100	10,722	13,147	6,211
		WGS	11,842	5,717	0	0

69 **Table 1. Study sample sizes.** The number of individuals of European ancestry included in the association analyses,
70 post-quality control (Methods). WES: whole-exome sequencing; WGS: whole-genome sequencing.

71
72 **CF-causing variant deltaF508 confers a protective effect against IBD.** The *CFTR* deltaF508
73 is an in-frame deletion variant commonly observed in the European population, with a minor
74 allele frequency (MAF) of 1.5% in non-Finnish Europeans (chr7:117559590:ATCT:A in
75 Genome Reference Consortium Human Build 38 [GRCh38]). Homozygous deltaF508 has full
76 penetrance for CF and is the most common CF-causing variant. Single variant association
77 analysis in the Broad discovery dataset using a logistic mixed model (Methods) found that
78 deltaF508 had a significant protective effect against both CD (p-value=1.53E-7, OR=0.73 [0.65-
79 0.82, 95% CI]) and UC (p-value=3.35E-4, OR=0.79 [0.69-0.90, 95% CI], Table 2). This
80 association also reached nominal significance for CD in the Sanger WES and WGS datasets (p-
81 value=3.52E-02 and 8.32E-3, Table 2) with both protective effects. After meta-analysis, the
82 protective effect of deltaF508 on CD reached genome-wide significance (p-value=5.5E-9). There
83 are no known IBD-associated variants (defined as variants with posterior inclusion probability >
84 5% from fine-mapping²⁰ or reported in the published sequencing and genome-wide association
85 studies²¹⁻²³) within 1 million base pairs of deltaF508, therefore it is highly unlikely that the
86 deltaF508 association tags a known IBD genetic association.

87

Study	Subtype	MAF (case)	MAF(control)	p-value	BETA	SE	OR (95% CI)
Broad WES (Discovery)	CD	0.0097	0.0136	1.53E-07	-0.305	0.059	0.65 - 0.82
	UC	0.0096	0.0136	3.35E-04	-0.234	0.066	0.69 - 0.90
	IBD	0.0097	0.0136	8.50E-10	-0.289	0.047	0.68 - 0.82
Sanger WES (Replication)	CD	0.0135	0.0161	3.52E-02	-0.136	0.064	0.77 - 0.99
	UC	0.0141	0.0161	2.04E-01	-0.073	0.058	0.83 - 1.04
	IBD	0.0137	0.0161	1.47E-02	-0.108	0.044	0.82 - 0.98
Sanger WGS (Replication)	CD/IBD	0.0114	0.0157	8.32E-03	-0.310	0.117	0.58 - 0.92
Meta-analysis	CD	-	-	5.49E-09	-0.238	0.041	0.73 - 0.85
	UC	-	-	9.95E-04	-0.143	0.043	0.79 - 0.94
	IBD	-	-	8.96E-11	-0.201	0.031	0.77 - 0.87

88 **Table 2. Association between deltaF508 and CD, UC, and IBD.** MAF: minor allele frequency; BETA, SE: effect
89 size, and standard error from a logistic mixed model; OR (95% CI): 95% confidence interval of odds ratio (OR).

90

91 To limit the potential for confounded CF status to bias the allele frequency estimates in our cases
92 and controls, we repeated the association analysis after excluding 4 cases and 47 controls who
93 were homozygous carriers of deltaF508 in the discovery and replication cohorts combined, as
94 well as 37 cases and 170 controls deemed potential compound heterozygous carriers
95 (Supplementary Tables 1), defined as individuals with two or more variants annotated as "CF-
96 causing" or "Varying clinical consequence" in the Clinical and Functional Translation of *CFTR*²⁴
97 (*CFTR2*) database. We found that the associations between *CFTR* deltaF508 and CD/UC/IBD
98 remain significant (Supplementary Tables 2), supporting the conclusion that deltaF508 plays a
99 protective role in IBD.

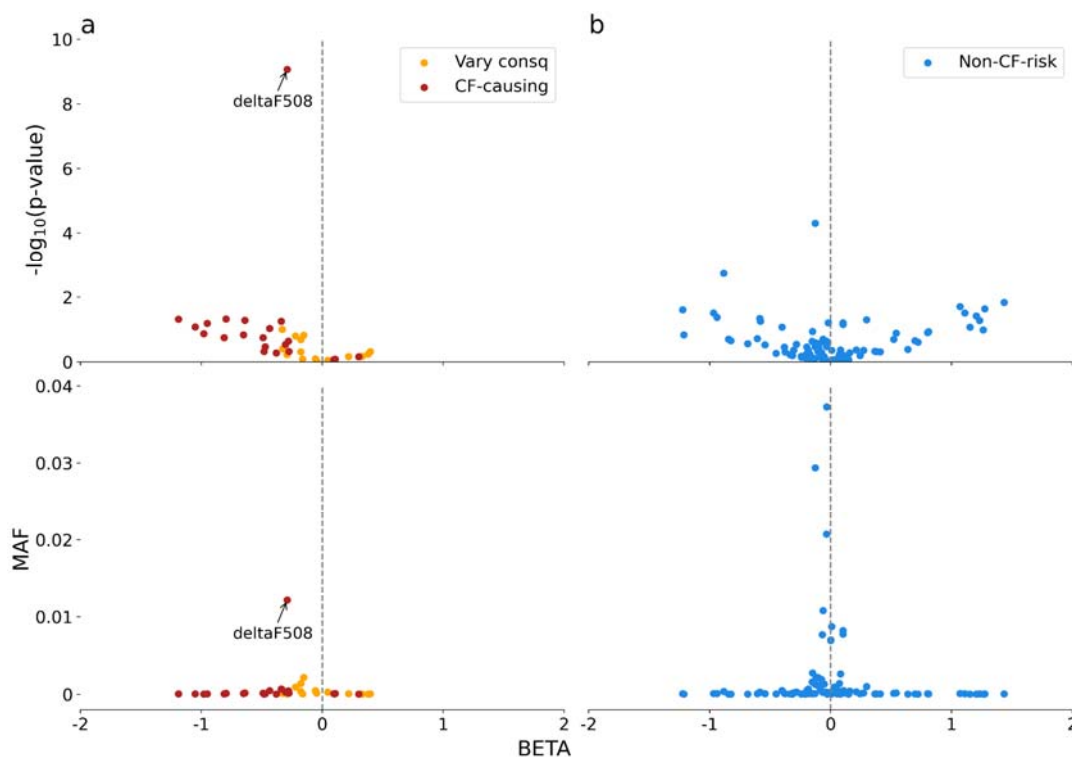
100

101 ***CFTR* mutations confer protection against IBD in gene-based burden tests.**

102 The *CFTR2* database contains a comprehensive list of *CFTR* variants known to impair *CFTR*
103 protein function and cause CF. The Broad discovery dataset captured 1036 *CFTR* variants
104 (Supplementary Tables 3). 170 of the 1035 *CFTR* variants were annotated in *CFTR2* as "CF-
105 causing" (109), "Varying clinical consequence" (33), "Non CF-causing" (19), or "Unknown
106 significance" (9). The remaining 865 variants did not have an annotation in *CFTR2*. We defined
107 "CF-risk" variants (142) as variants annotated as "CF-causing" or "Varying clinical
108 consequences" in the *CFTR2* database, and "Non-CF-risk" variants (893) as variants annotated
109 as "Non CF-causing", "Variants of unknown significance" in *CFTR2*, or without annotation in
110 *CFTR2*.

111

112 While no variants other than deltaF508 reached even nominal levels of significance (p-
113 value<0.01) in our single variant tests, CF-risk variants are predominantly protective against IBD
114 (p-value=0.002, binomial test using variants with minor allele count ≥ 10 in the Broad dataset,
115 Figure 1a). Conversely, non-CF-risk variants were not enriched with protective effects (p-
116 value=0.35, binomial test using variants with minor allele count ≥ 10 in the Broad dataset, Figure
117 1b).



118
 119 **Figure 1: Single-variant association tests for *CFTR* variants with minor allele count (MAC) ≥ 10 in the Broad**
 120 **dataset. a) Distribution of p-values (from logistic mixed model) and MAF for 37 CF-risk variants. b) Distribution**
 121 **of p-values (from logistic mixed model) and MAF for 96 non-CF-risk variants.**

122
 123 Based on these observations, we hypothesize that variants impairing *CFTR* function are
 124 protective against IBD. To test this hypothesis, we did an unweighted burden test of *CFTR* using
 125 the 141 CF-risk variants (deltaF508 not included as itself confers a significant protective effect).
 126 We found these variants in aggregation have a significant protective effect on IBD in the Broad
 127 dataset ($P=4.3E-06$, $\beta=-0.21$, $se=0.047$) (Table 3), which was replicated in the Sanger WES
 128 dataset at nominal significance ($P=0.037$, $\beta=-0.11$, $se=0.052$) and in the Sanger WGS dataset
 129 for the direction and effect size with CD ($p\text{-value}=0.073$, $\beta=-0.15$, $se=0.088$). The Sanger
 130 WGS is much smaller in the sample size thus, the lack of significance is likely due to its limited
 131 power. Combining all three datasets using the fixed-effect meta-analysis, we found CF-risk
 132 variants in *CFTR* collectively confer a protective effect of -0.164 ($OR=0.85$) with $p\text{-value}=3.9E-$
 133 7 . This association was significant when tested for CD and UC separately (meta-analysis for CD:
 134 $\beta=-0.13$, $se=0.039$, $p\text{-value}=8.8E-04$; UC: $\beta=-0.23$, $se=0.049$, $p\text{-value}=3.6E-06$)
 135 (Supplementary Table 4). On the contrary, we did not observe a significant association with IBD
 136 using the non-synonymous variants that were classified as non-CF-risk in *CFTR* ($p\text{-value}=0.37$
 137 and 0.56 , Table 3). Taken together, these results confirm our hypothesis that rare variants that
 138 impair *CFTR* function reduce the risk of IBD.

139

Study	Subtype	Molecular annotation	Clinical annotation	MAF	# of Variants	CAF	BETA	SE	p-value
-------	---------	----------------------	---------------------	-----	---------------	-----	------	----	---------

Broad WES	IBD	-	CF-risk	-	141	0.012	-0.21	0.047	4.3E-06
Sanger WES	IBD	-	CF-risk	-	128	0.011	-0.11	0.052	0.037
Sanger WGS	CD/IBD	-	CF-risk	-	56	0.023	-0.16	0.088	0.073
Meta-analysis		-	CF-risk	-	-	-	-0.164	0.032	3.9E-07
Broad WES	IBD	Non-synonymous	non-CF-risk	<0.001	697	0.017	-0.03	0.038	0.37
Sanger WES	IBD	Non-synonymous	non-CF-risk	<0.001	844	0.021	-0.063	0.037	0.090
Sanger WGS	CD/IBD	Non-synonymous	non-CF-risk	<0.001	180	0.020	-0.058	0.097	0.55

140 **Table 3. CFTR burden test results using clinical variant annotations.** CAF: composite allele frequency, defined
 141 as the frequency of observing carriers of at least one variant of interest in the study. We applied an MAF upper
 142 bound to burden tests using non-CF-risk variants so that the CAFs are similar to those of CF-risk variants.

143

144 **Variant annotation plays a critical role in burden tests**

145 The power of rare variant burden tests increases with the composite allele frequency and the
 146 proportion of causal to non-causal rare variants included in the analysis. Therefore, identifying a
 147 variant set that maximizes the inclusion of causal variants while minimizing non-causal variants
 148 is crucial to enhancing the power of the burden test. A common approach employed by the
 149 burden test is to use predicted loss of function variants or rare (MAF < 0.1%) non-synonymous
 150 variants. In both scenarios, we found much weaker evidence of association (p-value=1.01E-03
 151 and 0.042, respectively, Table 3) compared to the burden test using CF-risk variants. This
 152 demonstrates that clinical variant annotations from the CFTR2 database outperform naive
 153 molecular consequence annotations in detecting the impact of rare CFTR variants on IBD risk.

154

Molecular annotation	Number of Variants	CAF	BETA	SE	p-value
Predicted loss-of-function	93	0.003	-0.18	0.093	0.042
Non-synonymous	782	0.023	-0.10	0.032	1.01E-03

155 **Table 4. CFTR burden test using molecular annotations.** Definitions of “predicted loss-of-function” and “non-
 156 synonymous” variants are described in Methods. We restricted the tests to variants with MAF<0.1%.

157

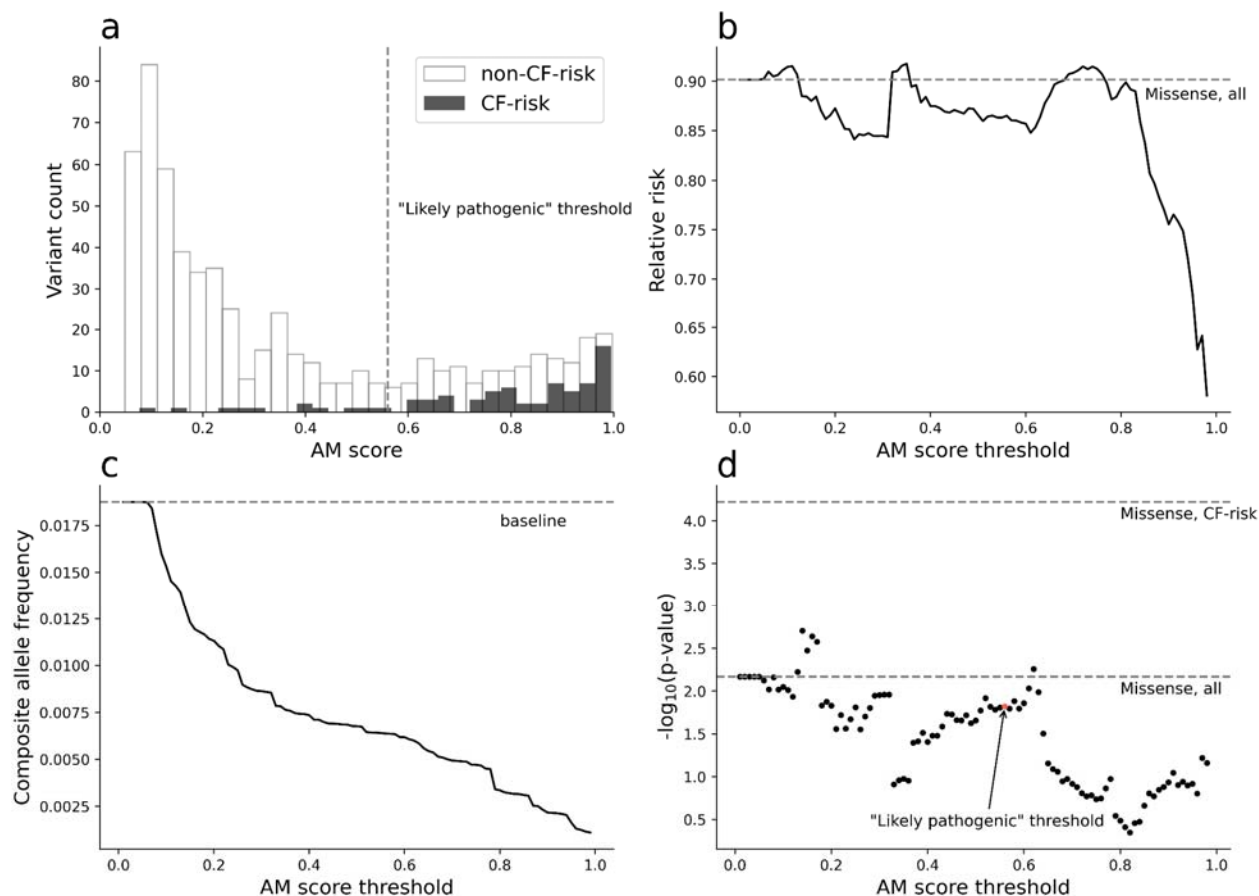
158 Unfortunately, very few genes are as thoroughly studied and clinically annotated as CFTR. Most
 159 human genes lack accurate variant effect annotation. *In silico* variant pathogenicity predictions
 160 can thus play an important role in systematically annotating variants to enhance the power of rare
 161 variant burden tests. Recently, Alpha Missense (AM)²⁵, a machine learning model trained on
 162 protein structure prediction and evolutionary constraint information for predicting missense
 163 variant pathogenicity, has been shown to have more accurate missense pathogenicity predictions
 164 as evidenced by better correlations with functional assays than previous prediction models²⁶.
 165 Thus, we evaluated the utility of AM for prioritizing CFTR missense variants for inclusion in
 166 burden tests, using the CFTR2 database as the benchmark.

167

168 We annotated 698 CFTR missense variants available in the Broad WES dataset, among which 74
 169 are CF-risk variants according to the CFTR2 database (Methods). We observed that most CF-risk

170 variants were predicted as “likely pathogenic” (AM score > 0.56) (Figure 2a), and variants with
171 high AM scores (> 0.8) had stronger protective effects against IBD (Figure 2b). However, 73% of
172 predicted “likely pathogenic” variants were non-CF-risk in CFTR2. This finding is concordant
173 with those from previous benchmarking studies that showed AM, as well as other pathogenicity
174 prediction models, tend to overcall pathogenic variants^{26,27}. The power to detect association via a
175 variant burden test could therefore be heavily influenced by the AM score threshold used to
176 define the set of variants included in the test. To investigate this, we performed burden tests by
177 selecting *CFTR* missense variants (restricted to MAF < 0.1%) with AM score threshold ranging
178 from 0 to 1 in increments of 0.01.

179
180 Our analysis showed that a higher AM score threshold does not always improve the statistical
181 power of the burden test. Statistical power is sensitive to both the effect size (increases with the
182 AM score threshold, Figure 2b) and the composite allele frequency (CAF, decreases with the
183 AM score threshold, Figure 2c). Specifically, within the score range of 0.56 to 1, the burden test
184 significance decreased with more stringent AM score cutoffs due to a reduction in CAF, despite
185 a stronger effect size (Figure 2d). To effectively incorporate pathogenicity scores in burden tests,
186 such as AM, further improvements on *in silico* variant annotations are needed to match the
187 accuracy of clinical annotations (Discussion).



188

189 **Figure 2: Prioritizing *CFTR* variants in burden tests using AlphaMissense.** **a)** AM scores for CF-risk variants
190 and non-CF-risk variants. **b)** Relative risk for missense variants with AM score above the AM threshold (x-axis).
191 Relative risk was calculated as CAF in cases / CAF in controls. **c)** CAFs for missense variants with scores above the
192 AM threshold (x-axis). **d)** Burden test significance using missense variants with AM scores above the threshold (x-
193 axis). The result using the AM “likely pathogenic” threshold (0.56) is marked with red. We restricted the tests to
194 variants with MAF<0.1%.

195

196 Discussion

197 Leveraging a large-scale IBD exome sequencing dataset, we showed at single-variant and gene-
198 based levels that CF-risk variants in the *CFTR* gene confer a protective effect against IBD. The
199 gut epithelial barrier plays a fundamental role in maintaining intestinal homeostasis and
200 protecting against IBD²⁸. Given *CFTR* mutations are associated with mucus buildup on the
201 epithelial cell surface, it is possible that heterozygous *CFTR* mutations alter the intestinal mucosa
202 which provides an enhancing effect on the gut epithelial barrier. Using a CF mouse model,
203 Gabriel *et al*²⁹ showed that heterozygote carriers of pathogenic *CFTR* variants are resistant to
204 cholera toxin. This resistance is due to reduced intestinal fluid and chloride ion secretion in
205 response to the toxin. In addition, it has been demonstrated that *CFTR* can serve as an epithelial
206 receptor for *S. Typhi* transluminal migration and that heterozygous deltaF508 mice translocated
207 significantly fewer *S. Typhi* into the gastrointestinal submucosa than wild-type *CFTR* mice³⁰.
208 Therefore, it is also plausible that the protective effect of *CFTR* in IBD may stem from similar
209 interactions with yet unidentified bacteria¹⁹. Our finding suggests a previously under-appreciated
210 role of CF-risk mutations in human health and disease. Follow-up experimental studies are
211 warranted to investigate the exact cellular pathways CF-risk mutations disrupt to exert such a
212 protective effect. This may provide insights for an effective therapeutic intervention for IBD.

213

214 The ascertainment bias in CF patients can potentially confound this study. The Broad study
215 ascertained control subjects from hospitals, while the Sanger study mostly ascertained controls
216 from the community via UK Biobank. Therefore, an over-representation of CF patients in the
217 Broad study can potentially create a spurious association between CF-risk variants and IBD. We
218 controlled for this confounding factor by removing all predicted CF patients based on genetic
219 data. While this approach is an approximation and does not guarantee the removal of all CF
220 patients, the remaining number of CF patients is very small and should not affect the validity of
221 our conclusion.

222

223 Burden tests are sensitive to the quality of variant annotations. While the clinical evidence-based
224 *CFTR* variant annotations established the protective effect of CF-risk variants on IBD, simply
225 including all predicted loss-of-function or missense variants in the burden test failed to reach
226 exome-wide statistical significance. We evaluated AM, the best-in-class *in-silico* variant
227 pathogenicity predictor, and found that its effectiveness in variant prioritization for burden tests
228 is hampered by insufficient accuracy. Further improvement in pathogenicity prediction models
229 for missense variants is needed. Furthermore, a higher score threshold for selecting variants for

230 burden tests does not guarantee improvement in statistical power as it may sacrifice sensitivity
231 for specificity, both of which are needed for a powerful burden test.

232
233 While this study focuses on the effect of *CFTR* mutations on IBD, the exome-wide analysis,
234 incorporating additional samples sequenced at Sanger Institute, is underway. In the very near
235 future, we expect to report a comprehensive list of IBD-associated genes through large-scale
236 exome sequencing analysis.

237

238 **ACKNOWLEDGMENTS**

239 We thank all of the principal investigators, local staff from individual cohorts, and all of the
240 patients who kindly donated samples used in the study for making possible this global
241 collaboration and resource to advance IBD genetics research. This research was funded in whole,
242 or in part, by the US National Institutes of Health grants no. U54HG003067 and no.
243 5UM1HG008895, the Wellcome Trust grants no. [206194] and no. [108413/A/15/D], and The
244 Leona M. & Harry B. Helmsley Charitable Trust grant no. 2015PG-IBD001. We thank the Broad
245 Institute Genomics Platform for genomic data generation efforts and the Stanley Center for
246 Psychiatric Research at the Broad Institute for supporting control sample aggregation. We thank
247 the Sanger Institute Scientific Operations teams and Human Genetics Informatics team for
248 sample handling and data generation. This research was supported by the NIHR IBD
249 BioResource and NIHR Biomedical Research Centres in Cambridge, Oxford, Imperial, UCH and
250 Newcastle. The views expressed are those of the authors and not necessarily those of the NIHR
251 or the Department of Health and Social Care. The NIHR IBD BioResource acknowledges co-
252 funding by Crohn's Colitis UK and The Leona M. and Harry B. Helmsley Charitable Trust. H.H.
253 acknowledges support from NIDDK grant no. R01DK129364 and the Stanley Center for
254 Psychiatric Research. Individual studies contributing to this project acknowledge support from
255 NIH grants no. DK062431, no. DK062432, no. DK087694, no. K23DK117054, no.
256 R01DK111843, no. P01DK094779, no. R01HG010140, no. 5U01HG009080 and no. DK062420,
257 and NIDDK grants no. P01DK046763, no. U01DK062413, DK 043351 and no. R01DK104844.

258

259 **AUTHOR CONTRIBUTIONS**

260 H.H., C.A.A. and M.J.D. designed and supervised the study. C.R.S., L.F., H.H., C.A.A. and
261 M.J.D. were responsible for project management. M.Y., Q.Z., K.Y., and A.S. performed data
262 analysis. M.Y. and H.H. wrote the manuscript. All authors have reviewed and approved the
263 manuscript.

264

265 **COMPETING INTERESTS**

266 M.J.D. is a founder of Maze Therapeutics. C.A.A. has received consultancy fees from Genomics
267 plc and BridgeBio Inc. and lecture fees from GSK. The remaining authors declare no competing
268 interests.

269 **METHODS**

270 **Ethics.** All relevant ethical guidelines have been followed, and any necessary institutional
271 review board (IRB) and/or ethics committee approvals have been obtained. The Broad Institute
272 component of this study was approved under Study Protocol 2013P002634 (the Broad Institute
273 Study of Inflammatory Bowel Disease Genetics), and undergoes annual continuing review by the
274 Mass General Brigham Human Research Committee IRB of Mass General Brigham (Mass
275 General Brigham IRB). Approval was given on 27 January 2021 for this study. DNA samples
276 sequenced at the Sanger Institute were ascertained under the following ethical approvals:
277 12/EE/0482, 12/YH/0172, 16/YH/0247, 09/H1204/30, 17/EE/0265, 16/WM/0152, 09/H0504/125,
278 15/EE/0286, 11/YH/0020, 09/H0717/4, REC 22/02, 03/5/012, 03/5/012, 2000/4/192,
279 05/Q1407/274, 05/Q0502/127, 08/ H0802/147, LREC/2002/6/18, GREC/03/0273 and
280 YREC/P12/0. All informed consent from participants has been obtained, and the appropriate
281 institutional forms have been archived.

282

283 **Broad Institute data production**

284 *Sequencing.* The sequencing process included sample preparation (Illumina Nextera, Illumina
285 TruSeq, and Kapa Hyperprep), hybrid capture (Illumina Rapid Capture Enrichment - 37 Mb
286 target [“Nextera”, Table 1], and Twist Custom Capture - 37 Mb target [“Twist”, Table 1]) and
287 sequencing (Illumina HiSeq2000, Illumina HiSeq2500, Illumina HiSeq4000, Illumina HiSeqX,
288 Illumina NovaSeq 6000, 76-base pair (bp) and 150-bp paired reads). Sequencing was performed
289 at a median depth of 85% targeted bases at $>20\times$. Sequencing reads were mapped by BWA-
290 MEM to the hg38 reference using the GATK ‘functional equivalence’ pipeline. The mapped
291 reads were then marked for duplicates, and base quality scores were recalibrated. They were then
292 converted to CRAMs using Picard 2.16.0-SNAPSHOT and GATK 4.0.11.0. The CRAMs were
293 then further compressed using ref-blocking to generate gVCFs. These CRAMs and gVCFs were
294 then used as inputs for joint calling. To perform joint calling, the single-sample gVCFs were
295 hierarchically merged.

296

297 *Quality control.* Quality control (QC) analyses were conducted in Hail v.0.2.128 (Supplementary
298 Figure 1). We first split multiallelic sites and coded genotypes with low genotype quality ($GQ <$
299 20) as missing. To exclude variant sites that fall outside of exome capture, we removed variants
300 that are not annotated as frameshift, inframe deletion, inframe insertion, stop lost, stop gained,
301 start lost, splice acceptor, splice donor, splice region, missense or synonymous. Sample-level QC.
302 Samples that satisfy the following conditions were removed: (1) samples with an extremely large
303 number of singletons (≥ 500); (2) samples with mean $GQ < 30$; (3) samples with missingness
304 rates $> 10\%$; (4) samples with outlying heterozygosity (± 5 s.d. away from mean within the
305 population); (5) samples with inconsistent genetically imputed sex and reported sex; and (6)
306 duplicated samples, which were removed by identifying pairs of samples sharing identical
307 genotypes ($PI_HAT > 0.95$) and keeping the sample with higher mean GQ . Variant-level QC.
308 Variants that satisfy following conditions were removed: (1) variants with missingness rate $> 5\%$;

309 (2) variants with mean read depth (DP) < 10; (3) variants with >10% samples that were
310 heterozygous and with an allelic balance ratio <0.3 or >0.7; and (4) variants that have known
311 quality issues in both gnomAD v2 and v3 dataset (non-empty values in the “filter” column).

312
313 *Ancestry assignment.* We trained a random forest classifier using 1000 Genomes Project (1KGP)
314 subjects and ten principal components (PC) derived from a set of ~22000 common variants
315 shared between 1KGP and our callset. We projected all our samples onto the PC space generated
316 based on the 1KGP subjects and assigned each of our subject to European (CEU, TSI, FIN, GBR,
317 IBS), African and American (YRI, LWK, GWD, MSL, ESN, ASW, ACB, MXL, PUR, CLM,
318 PEL), East Asian (CHB, JPT, CHS, CDX, KHV), and South Asian (GIH, PJL, BEB, STU, ITU).
319 For this study, we kept samples that were classified as European with a prediction probability
320 greater than 80% (Supplementary Figure 2).

321
322 **Sanger Institute data production**

323 *Sequencing.* Genome sequencing was performed at the Sanger Institute using the Illumina
324 HiSeqX platform with a combination of PCR and PCR-free library preparation protocols.
325 Sequencing was performed at a median depth of 18.6x. Exome sequencing of IBD cases was
326 performed at the Sanger Institute using the Illumina NovaSeq 6000 and the Agilent SureSelect
327 Human All Exon V5 capture set. Controls from the UK Biobank were sequenced separately as a
328 part of the UKBB WES200K release using Illumina NovaSeq and the IDT xGen Exome
329 Research Panel v1.0 capture set (including supplemental probes). 168,100 UKBB participants
330 with EUR ancestry were selected for use as controls, excluding participants with recorded or
331 self-reported CD, UC, unspecified noninfective gastroenteritis or colitis, any other immune-
332 mediated disorders, or a history of being prescribed any drugs used to treat IBD. Reads were
333 mapped to hg38 reference using BWA-MEM 0.7.17. Variant calls were performed using
334 DeepVariant and saved as per-sample gVCFs. These gVCFs are aggregated with GLnexus into
335 joint-genotyped, multi-sample project-level VCFs (pVCFs). Variant calling was limited to
336 Agilent extended target regions. Per-region VCF shards were imported into the Hail software and
337 combined. This study considered variants located in intersection regions of Agilent and IDT
338 exome captures + 100bp buffer.

339
340 *Quality control* Exome sequencing: A combination of filters was used to identify low-quality
341 variants and samples. Genotype calls with low genotype quality (GQ < 20) in rare variants (MAF
342 < 0.1%) were set as missing. A variant level QC was then applied by keeping variants that meet
343 all the following criteria in both case (Sanger WES) and control (UKBB WES) samples: (1)
344 mean GQ > 30; (2) mean read depth (DP) > 10; and (3) call rate > 0.95. Samples satisfying any
345 of the following conditions were removed: (1) low average GQ (≤ 30); (2) low call rate (≤ 0.9);
346 (3) disagreement between genetically predicted and reported sex; (4) genetically identified
347 duplicates (samples with higher call rate were retained); and (5) high rates of heterozygosity (± 4
348 s.d. away from mean within each dataset). Genome sequencing: we applied variant quality score

349 recalibration (VQSR) to calculate the variant quality score log-odds (VQSLOD) for each variant
350 using GATK v4.4. Variants in the range of VQSLODs corresponding to the remaining 0.5% of
351 the truth set were removed. Furthermore, we kept variants that meet all the following criteria: (1)
352 mean GQ > 30; (2) mean read depth (DP) > 10; and (3) call rate > 0.9. Details on sample QC
353 were available elsewhere²².

354
355 *Ancestry assignment* We selected a set of ~14,000 high-quality common variants that were
356 shared between our subjects and 1KGP subjects for ancestry assignment. Using this set of
357 variants, we created four principal components from the 1KGP subjects and projected our
358 subjects to these components. We then used Random Forest to classify samples into broad
359 genetic ancestry groups (EUR, AFR, SAS, EAS, admixed), with 1KGP as the training dataset.
360 We only retained the EUR samples for this study, as the number of cases for other ancestry
361 groups was too small for robust association analysis.

362
363 **Association analysis. Broad Institute:** Association analyses were performed using a logistic
364 mixed model implemented in REGENIE v.2.2.4. A set of high-confidence variants (>1% MAF,
365 99% call rate, LD-pruned) was used for polygenic effect parameter estimation (Step 1). To
366 control for case-control imbalance, Firth correction was applied to association tests with p-value
367 < 0.05. To control for residual population structure, we calculated five PCs using a set of well-
368 genotyped common SNPs, excluding regions with known long-range LD. We included a binary
369 variable representing the sample sequencing platform to control for the sequencing heterogeneity
370 (Twist or Nextera). These variables were used as covariates in both single-variant and burden
371 tests, along with sex and the polygenic effect parameter calculated in Step 1. **Sanger Institute:**
372 A similar pipeline has been applied at the Sanger Institute. We used REGENIE v3.1.2 to
373 calculate the leave-one-chromosome-out (LOCO) score (Step 1) based on a set of high-quality
374 variants (MAF > 1%, HWE P > 1e-15, LD-pruned with R²=0.9, and not from long-range LD
375 regions). We then performed logistic regression using the LOCO score, sex, and four PCs as
376 covariates in the model (Step 2). PCs were calculated based on variants with MAF > 0.1% and
377 HWE P > 1e-15; variants from long-range LD regions or known IBD regions were excluded.
378 We applied the fast Firth correction (flags: "--firth --approx") to association tests with p-values <
379 0.05 to correct for effect size estimation bias caused by case-control imbalance.

380
381 **Meta-analysis.** We used METAL³¹ with an IVW fixed-effect model to meta-analyze
382 the association statistics across different studies.

383
384 **Clinical variant Annotation.** Cystic Fibrosis variant annotation was downloaded from the
385 Variant List History tab on CFTR2.org on April 23rd, 2024. Each variant was mapped to variant
386 ID in GRCh38 by its cDNA name.

387

388 **Variant annotation for molecular effect.** Non-synonymous variants were classified using
389 Ensembl Variant Effect Predictor (VEP v.95.0 in Broad institute and VEP v.110.1 in Sanger
390 institute)³² as one of the following most severe consequences: "frameshift_variant",
391 "stop_gained", "splice_acceptor_variant", "splice_donor_variant", "inframe_deletion",
392 "inframe_insertion", "stop_lost", "start_lost", "missense_variant". Predicted loss-of-function
393 variants are defined as variants annotated by VEP as one of the following mutation types:
394 "frameshift_variant", "stop_gained", "splice_acceptor_variant", "splice_donor_variant".

395
396 **Variant Annotation using AlphaMissense.** AlphaMissense (AM) predictions for all single
397 amino acid substitutions in the human proteome data were downloaded and subsetted to 698
398 CFTR missense variants available in the Broad discovery dataset.

399
400 **DATA AVAILABILITY**
401 Genome Reference Consortium Human Build 38 can be accessed at
402 <https://www.ncbi.nlm.nih.gov/assembly/>. Cystic Fibrosis variant annotation data can be
403 downloaded from the Variant List History tab on CFTR2.org. AlphaMissense pathogenicity
404 scores can be downloaded from <https://zenodo.org/records/8208688>. Sequence data used in this
405 study has been made publicly available in dbGaP Study Accession: phs001642.v2.p1 - Center for
406 Common Disease Genomics [CCDG] - Autoimmune: Inflammatory Bowel Disease (IBD)
407 Exomes and Genomes ([https://www.ncbi.nlm.nih.gov/projects/gap/cgi-](https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001642.v2.p1)
408 [bin/study.cgi?study_id=phs001642.v2.p1](https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001642.v2.p1)).

409
410 **CODE AVAILABILITY**
411 The software and code used are described throughout the Methods and can be found at
412 <https://github.com/mingRYU/CFTR-IBD>.

413
414

415 **Reference**

- 416 1. Welsh, M. J. & Smith, A. E. Molecular mechanisms of CFTR chloride channel dysfunction
417 in cystic fibrosis. *Cell* **73**, 1251–1254 (1993).
- 418 2. Anderson, M. P. *et al.* Demonstration that CFTR is a chloride channel by alteration of its
419 anion selectivity. *Science* **253**, 202–205 (1991).
- 420 3. Cheng, S. H. *et al.* Defective intracellular transport and processing of CFTR is the
421 molecular basis of most cystic fibrosis. *Cell* **63**, 827–834 (1990).
- 422 4. Lara-Reyna, S., Holbrook, J., Jarosz-Griffiths, H. H., Peckham, D. & McDermott, M. F.
423 Dysregulated signalling pathways in innate immune cells with cystic fibrosis mutations.
424 *Cell. Mol. Life Sci.* **77**, 4485–4503 (2020).
- 425 5. Keiser, N. W. *et al.* Defective innate immunity and hyperinflammation in newborn cystic
426 fibrosis transmembrane conductance regulator-knockout ferret lungs. *Am. J. Respir. Cell*
427 *Mol. Biol.* **52**, 683–694 (2015).
- 428 6. Rottner, M., Kunzelmann, C., Mergey, M., Freyssinet, J.-M. & Martínez, M. C.
429 Exaggerated apoptosis and NF-kappaB activation in pancreatic and tracheal cystic fibrosis
430 cells. *FASEB J.* **21**, 2939–2948 (2007).
- 431 7. Venkatakrisnan, A. *et al.* Exaggerated activation of nuclear factor-kappaB and altered
432 IkappaB-beta processing in cystic fibrosis bronchial epithelial cells. *Am. J. Respir. Cell*
433 *Mol. Biol.* **23**, 396–403 (2000).
- 434 8. Zhang, Y.-L. *et al.* Increased intracellular Cl concentration promotes ongoing inflammation
435 in airway epithelium. *Mucosal Immunol.* **11**, 1149–1157 (2018).
- 436 9. Bruscia, E. M. & Bonfield, T. L. Innate and Adaptive Immunity in Cystic Fibrosis. *Clin.*
437 *Chest Med.* **37**, 17–29 (2016).

- 438 10. Simonin-Le Jeune, K. *et al.* Impaired functions of macrophage from cystic fibrosis patients:
439 CD11b, TLR-5 decrease and sCD14, inflammatory cytokines increase. *PLoS One* **8**, e75667
440 (2013).
- 441 11. Ralhan, A. *et al.* Current Concepts and Controversies in Innate Immunity of Cystic Fibrosis
442 Lung Disease. *J. Innate Immun.* **8**, 531–540 (2016).
- 443 12. Ratner, D. & Mueller, C. Immune responses in cystic fibrosis: are they intrinsically
444 defective? *Am. J. Respir. Cell Mol. Biol.* **46**, 715–722 (2012).
- 445 13. Aldallal, N. *et al.* Inflammatory response in airway epithelial cells isolated from patients
446 with cystic fibrosis. *Am. J. Respir. Crit. Care Med.* **166**, 1248–1256 (2002).
- 447 14. Becker, M. N. *et al.* Cytokine secretion by cystic fibrosis airway epithelial cells. *Am. J.*
448 *Respir. Crit. Care Med.* **169**, 645–653 (2004).
- 449 15. Bolton, C. *et al.* An Integrated Taxonomy for Monogenic Inflammatory Bowel Disease.
450 *Gastroenterology* **162**, 859–876 (2022).
- 451 16. Bresso, F. *et al.* Potential role for the common cystic fibrosis DeltaF508 mutation in
452 Crohn’s disease. *Inflamm. Bowel Dis.* **13**, 531–536 (2007).
- 453 17. Bahmanyar, S., Ekbom, A., Askling, J., Johannesson, M. & Montgomery, S. M. Cystic
454 fibrosis gene mutations and gastrointestinal diseases. *J. Cyst. Fibros.* **9**, 288–291 (2010).
- 455 18. Shakhnovich, V. *et al.* P-199 pathogenic CFTR mutation in crohn’s disease in the absence
456 of other CFTR-related manifestations. *Inflamm. Bowel Dis.* **22**, S69 (2016).
- 457 19. Bresso, F. & D’Amato, M. The cystic fibrosis F508del mutation in Crohn’s disease. *J. Cyst.*
458 *Fibros.* **10**, 132 (2011).
- 459 20. Huang, H. *et al.* Fine-mapping inflammatory bowel disease loci to single-variant resolution.
460 *Nature* **547**, 173–178 (2017).

- 461 21. Rivas, M. A. *et al.* Deep resequencing of GWAS loci identifies independent rare variants
462 associated with inflammatory bowel disease. *Nat. Genet.* **43**, 1066–1073 (2011).
- 463 22. Sazonovs, A. *et al.* Large-scale sequencing identifies multiple genes and rare variants
464 associated with Crohn’s disease susceptibility. *Nat. Genet.* **54**, 1275–1283 (2022).
- 465 23. Liu, Z. *et al.* Genetic architecture of the inflammatory bowel diseases across East Asian and
466 European ancestries. *Nat. Genet.* **55**, 796–806 (2023).
- 467 24. The Clinical and Functional TRanslation of CFTR (CFTR2); available at <http://cftr2.org>.
- 468 25. Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with
469 AlphaMissense. *Science* **381**, eadg7492 (2023).
- 470 26. Ljungdahl, A. *et al.* AlphaMissense is better correlated with functional assays of missense
471 impact than earlier prediction algorithms. *bioRxiv* (2023) doi:10.1101/2023.10.24.562294.
- 472 27. McDonald, E. F., Oliver, K. E., Schleich, J. P., Meiler, J. & Plate, L. Benchmarking
473 AlphaMissense Pathogenicity Predictions Against Cystic Fibrosis Variants. *bioRxiv* (2023)
474 doi:10.1101/2023.10.05.561147.
- 475 28. Laukoetter, M. G., Nava, P. & Nusrat, A. Role of the intestinal barrier in inflammatory
476 bowel disease. *World J Gastroenterol* **14**, 401–407 (2008).
- 477 29. Gabriel, S. E., Brigman, K. N., Koller, B. H., Boucher, R. C. & Stutts, M. J. Cystic fibrosis
478 heterozygote resistance to cholera toxin in the cystic fibrosis mouse model. *Science* **266**,
479 107–109 (1994).
- 480 30. Pier, G. B. *et al.* *Salmonella typhi* uses CFTR to enter intestinal epithelial cells. *Nature* **393**,
481 79–82 (1998).
- 482 31. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of
483 genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).

484 32. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).