

1 Diversity of *CFTR* variants across ancestries characterized using 454,727 UK 2 biobank whole exome sequences

3
4 Justin E. Ideozu¹, Mengzhen Liu², Bridget M. Riley-Gillis², Sri R. Paladugu², Fedik Rahimov²,
5 Preethi Krishnan³, Rakesh Tripathi³, Patrick Dorr³, Hara Levy⁴, Ashvani Singh⁵, Jeffery F.
6 Waring^{1,2}, and Aparna Vasanthakumar¹.

7
8 ¹Genomic Medicine, Genomics Research Center, AbbVie, IL.

9 ²Human Genetics, Genomics Research Center, AbbVie, IL.

10 ³Precision Medicine, AbbVie, IL.

11 ⁴Department of Pediatrics, Division of Pulmonology and Sleep Medicine, University of
12 Wisconsin School of Medicine and Public Health, WI.

13 ⁵Discovery Research, AbbVie, IL.

14

15 **Keywords:** CFTR, whole exome sequencing, UK biobank, cystic fibrosis

16

17 Abstract

18

19 Limited understanding of the diversity of *CFTR* variants across ancestries hampers efforts to
20 advance molecular diagnosis of cystic fibrosis (CF). The consequences pose a risk of delayed
21 diagnoses and subsequently worsened health outcomes for patients. Characterizing the
22 distribution of *CFTR* variants across ancestries is, therefore, critical for revolutionizing
23 molecular diagnoses of CF. By interrogating 454,727 UK biobank whole exome Sequences, we
24 detected over 4,000 *CFTR* variants, including novel ancestry-specific variants, across six
25 ancestries. F508del was the most prevalent CF-causing variant found across the ancestries,
26 except in East Asia, where V520F was the most prevalent. Sixteen participants had two CF-
27 causing variants, with two diagnosed as CF. We found 154 participants with a CF-causing and
28 varying clinical consequences (VCC) variant. Overall, participants with multiple clinically
29 relevant variants reported indications significantly associated with CF and its pulmonary
30 phenotypes [Bonferroni adjusted $p < 0.05$]. Potentially, participants with classical CF phenotypes
31 may benefit from current CF therapies.

32

33 Introduction

34

35 In cystic fibrosis (CF), pathogenic variants in the cystic fibrosis transmembrane conductance
36 regulator (*CFTR*) gene results in system-wide debilitating consequences¹. Current *CFTR*
37 modulator therapies targeting the basic *CFTR* defects have yielded both short- and long-term
38 clinical benefits, but they are approved only for patients with certain *CFTR* variants^{2,3}. Even
39 within these patients, the response to current *CFTR* modulator therapies can be variable. Recent
40 advances in genetic technologies have led to the discovery of many CF-causing variants among
41 populations of European ancestry. Although these findings have led to a better understanding of
42 the disease prevalence, and improved diagnoses⁴, there is limited knowledge about the
43 distribution of ancestry-specific *CFTR* variants. Characterizing the broad spectrum of pathogenic
44 and non-pathogenic *CFTR* variants across ancestries holds promise to revolutionize molecular
45 diagnoses of CF and could enable wider access to personalized *CFTR* modulator therapies.

46

47 Most studies that have surveyed the diversity of *CFTR* variants have largely focused on studying
48 persons with CF from European populations^{5,6}. This is perhaps driven by the fact that CF
49 predominantly affects individuals from European ancestry, and the disease is less frequent in
50 people from other ancestries. Survey reports from a diverse US population shows CF affects
51 individuals of European ancestry (1:3000 live birth) relative to people from other ethnicities,
52 including Africans (1:15000), Asians (1:35,000), and Native Americans (1:10,900)^{7,8}. Globally,
53 CF is presumed to affect about 100,000 people and F508del accounts for most of the cases^{6,9}.
54 Recent reports suggest the spectrum of CF-causing variants vary across ancestries and
55 countries^{7,10}. CF-causing variants commonly reported in other ethnic groups, include
56 3120+1G>A (Africans)¹⁰, p.W1282X (Ashkenazim)^{6,11}, and p.G970D (Chinese)¹². Although
57 these studies have shed insights on common CF-causing variants in non-European populations,
58 carrier screening panels do not often capture the dominant CF-causing variants in some
59 ethnicities¹³. Thus, the prevalence of the disease in other ethnicities, and worldwide, are possibly
60 underestimated.

61
62 Diagnosis of CF can be based on the results of a patients' pilocarpine iontophoresis sweat test
63 (sweat chloride \geq 60 mmol/L) and/or molecular genetic testing (positive for up to two CF-
64 causing variants)¹⁴. One of the advantages of molecular genetic testing is that it could detect both
65 symptomatic and pre-symptomatic individuals, including the potential to determine a patients'
66 eligibility for current *CFTR* modulator therapies. A major drawback of solo reliance on
67 symptomatic diagnoses is that often patients are diagnosed when they are already presenting
68 symptoms, and delayed diagnosis negatively impacts the patients' health outcomes¹⁵. Early
69 diagnosis and interventions, particularly with triple *CFTR* therapy, results in improved clinical
70 outcomes and could profoundly affect the trajectory of CF lung disease¹⁶. Carriers of rare
71 pathogenic variants not included in carrier screening panels are more likely to have delayed
72 diagnosis which poses a risk of increased morbidity. Thus, it has become critical to characterize
73 the broad spectrum of pathogenic *CFTR* variants across ancestral populations.

74
75 Determining the clinical and functional relevance of *CFTR* variants is such an inundated task that
76 requires critical examination of the variants' impact on an individual's phenotype. This is
77 compounded by the clinical heterogeneity often seen in patients with the same *CFTR* genotype.
78 Efforts driven by CFTR1/2 have led to the identification of 2,110 *CFTR* variants
79 [<http://www.genet.sickkids.on.ca>]. Of these, 401 are CF-causing while 49 are variants of varying
80 clinical consequences (VCC), according to the latest CFTR2 annotation [<https://cftr2.org>]. The
81 CFTR1/2 databases have led to an increased understanding of CF diagnoses and prevalence
82 across populations. However, the variants captured from the CFTR1/2 efforts do not necessarily
83 represent the entirety of potential variants that could possibly exist within the *CFTR* gene across
84 ancestries. Recent annotations from TopMed Whole Genome Sequencing (WGS) efforts that
85 includes diverse populations indicate that the *CFTR* gene is about 250kb with 56,488 variants¹⁷,
86 which is longer than the often referenced 189kb length¹⁸. Although the vast majority of these
87 newly detected *CFTR* variants are intronic [96%], and their clinical relevance unknown, this
88 underscores the continued need to gain a comprehensive understanding of all pathogenic and
89 non-pathogenic variants that possibly exists within the *CFTR* gene, including those from various
90 ancestries.

91

92 The complexities associated with characterizing the clinical relevance of *CFTR* variants can be
93 subdued by leveraging population databases, such as UK Biobank (UKBB), that hosts both
94 genomics and phenotypic information of individuals. In this study, we characterized the diversity
95 of *CFTR* variants across six ancestral populations included in the UKBB. We segregated
96 ancestry-specific *CFTR* variants, including those that are CF-causing from uncharacterized
97 variants. We estimated the age of certain detected CF-causing variants. For participants with two
98 CF-causing variants or a CF-causing variant and a VCC, we performed phenotype analysis to
99 determine the possible impact of pathogenic CF-variants on their health outcomes. Our work
100 provides the foundation for future studies to explore the clinical relevance of the newly detected
101 *CFTR* variants across the described ancestries.

102

103 **Methods**

104

105 **Study population and whole exome sequences**

106

107 The dataset used for this study is available in the UKBB public repository and was accessed
108 under application 26041. A total of 454,787 UKBB whole exome sequences were interrogated to
109 characterize the diversity of *CFTR* variants across ancestries. The demographics and
110 characteristics of the UKBB participants has been described elsewhere¹⁹, but relevant
111 information such as gender, age and spirometry measures were recorded for each participant.
112 Details on the calling and quality control can be found in previous publication¹⁹. To maximize
113 variant discovery, no further filters were applied except the removal of individuals that
114 subsequently withdrew consent (final N = 454,727). Curated disease phenotypes, defined based
115 on participants primary care and hospital in-patient (HESIN) records, were utilized for
116 association tests with pathogenic *CFTR* genotypes.

117

118 **Ancestry designation**

119

120 For ancestry designation, we opted to use the classification from the PanUKBB working group
121 (<https://pan-dev.ukbb.broadinstitute.org/>) which derives the ancestry classification from two
122 large reference datasets, 1000 Genomes Project and Human Genome Diversity Project. Using
123 this classification, the participants were grouped into six major groups: African (AFR),
124 American/American Admixed (AMR), Central South Asia (CSA), East Asian (EAS), European
125 (EUR), and Middle East (MID). Individuals that did not fall neatly into the large continental
126 classifications were classified as MIX. While this classification limits the full extent of genomic
127 diversity in the UKBB, it permitted an estimation of variant frequencies to other populations.
128 Individuals assigned to the MIX group were included in all other counts and analysis.

129

130 **Variant annotation**

131

132 Clinically relevant variants (CF-causing and VCCs) detected across the ancestries were classified
133 using CFTR2 annotation. CF-causing variants without an assigned Reference SNP cluster ID
134 (rsID) were excluded since they comprised mostly of large indels and structural variants, which
135 are often challenging to be called by variant calling tools. Functionally relevant variants (high
136 impact variants) were annotated with Variant Effect Predictor²⁰ and SnpEFF²¹ and those high
137 impact variants not known to cause CF were recorded. The canonical transcript of *CFTR*

138 (ENST00000003084) was used as reference for both tools. CF is an autosomal recessive disease
139 and molecular diagnosis could be based on the possession of two or more clinically relevant
140 variants. Thus, participants with two CF-causing variants, a CF-causing variant and a VCC, and
141 a CF-causing variant and a non-CF-causing high impact variant, were characterized and
142 prioritized for further association test. Participants with two CF-causing variants were then
143 assigned likely disease severity status (pancreatic sufficient or insufficient) based on their *CFTR*
144 genotype as previously described^{22,23}.

145

146 **Statistical association analysis**

147

148 We categorized participants with certain clinical or functionally relevant variant combinations
149 into three groups: two CF-causing variants, a CF-causing variant and a VCC, a CF-causing
150 variant and a variant of high impact. We tested each of the groups for association with disease
151 phenotypes using Fisher-Exact tests. As pulmonary measures such as forced expiratory volume
152 (FEV) and forced vital capacity (FVC) could reflect pulmonary phenotype, we tested for
153 differences in these measures between participants with two CF-causing variants and the general
154 population. Only associations meeting a Bonferroni corrected adjusted p-value < 5% were
155 considered as statistically significant.

156

157 **Genealogical estimation of CF-causing variants age**

158

159 To estimate the age of CF-causing variants, we used data from the Atlas of Variant Age (AVA)
160 (<https://human.genome.dating/>). Models using data from the 1000 Genomes Project (TGP) and
161 the Simons Genome Diversity Project (SGDP), implemented in the Genealogical Estimation of
162 Variant Age (GEVA) tool²⁴, were deployed. For variant age estimations, joint mutational and
163 recombinational clocks were considered using an average generation time of 25 years. Estimated
164 age of variants within the *CFTR* gene locus (Chr7: 117,095,837 – 117,366,025) were then
165 reported.

166

167 **Result**

168

169 **Over 4000 *CFTR* variants identified across ancestries.**

170

171 We interrogated 454,787 whole exome sequences available in the UKBB to characterize the
172 diversity of *CFTR* variants across six ancestries [AFR, CSA, EAS, EUR, AMR, and MID] and
173 an uncharacterized group (MIX). The median age across all ancestries was greater than 60 years.
174 Except for CSA and MID where over half of the characterized participants were males, most of
175 the ancestries were dominated by females (Table 1). Overall, we detected 4,193 variants across
176 all ancestries in the UKBB, with V470M emerging as the most common variant (Table S1). The
177 *CFTR* variants represented a diverse range of variant types, but over 50% were intronic (Fig. 1)
178 The highest number of *CFTR* variants were detected in Europeans [n=3,192] while the AMR
179 group had the least number of *CFTR* variants [n=151]. Across the remaining ancestries [CSA,
180 AFR, EAS, and MID], the number of *CFTR* variants detected [471, 417, 266, and 222
181 respectively] were also much lower when compared to those detected in Europeans (Table 1).
182 Many of the variants detected in this study have never been reported as variants in *CFTR*. Next,
183 we explored if there were *CFTR* variants uniquely represented across the ancestries.

184 Interestingly, we found several *CFTR* variants that tend to be ancestry specific (Fig. 2, Table S2).
185 The highest number of unique *CFTR* variants [2212/3192] was found in Europeans, while the
186 Hispanic/Admixed American group had the least number of unique *CFTR* variants [23/151].
187 About 8.3% [37850/454,727] of the overall study population did not fall under a specific
188 ancestry and were categorized as the Mixed group (MIX). The median (interquartile range [IQR])
189 age for the MIX population was 67 (60, 67), with males accounting for 43% of the population.
190 Within the MIX group, we detected a total of 1333 *CFTR* variants and 28.6% of these variants
191 [n=381] were uniquely found in their population (Table 1).

192

193 **CF-causing variants across ancestries**

194

195 We explored the global distribution of CF-causing variants, variants of varying clinical
196 consequences (VCC), and high impact variants. As shown in Table 2, we found varying
197 distribution of CF-causing (Table S3) and VCC (Table S4) variants across the ancestries. The
198 highest number of CF-causing variants were detected in the EUR population [116/154]. These
199 variants accounted for 75% of all CF-causing variants with RSIDs captured in the study. Nearly
200 50% of the CF-causing variants were found in the MIX population, but lower proportions
201 (<15%) were reported across the other specific ancestries. The AMR group with the least number
202 of detected *CFTR* variants [151/4193] also had the least number [7/154] of the CF-causing
203 variants (Table 2). F508del was the most prevalent CF-causing variant observed in all ancestries,
204 except in EAS where V520F was the most prevalent (Fig. 3). While F508del represented ~90%
205 of the CF-causing variants found in Europeans, our findings indicate its prevalence was not as
206 high in other ancestries. We further explored whether there were any CF-causing variants that
207 were potentially ancestry-specific. As shown in Figure 4, all ancestries had unique CF-causing
208 variants, but EUR possessed the highest number [n=50] of ancestry-specific CF-causing variants
209 while four groups (AMR, CSA, EAS, and MID) had only one ancestry-specific variant each
210 (Table S5). We also detected VCC variants in all ancestries [Table S4]. The number of VCC
211 variants detected were smaller than the number of CF-causing variants detected across the
212 ancestries, except for AMR and MID populations. R74W was one of the most common variants
213 detected across the ancestries. Except for EUR, R74W featured among the top 5 most common
214 VCC reported in all ancestries. Other VCCs that made the common list included R117H,
215 D1270N, and L967S (Table S6). Interestingly, we found several high impact variants that are not
216 known to be CF-causing or clinically relevant in each ancestry. The highest number of these
217 variants [150/200] were detected in the EUR population. Overall, high impact variants accounted
218 for only a small fraction of the total number of variants reported in each ancestry [$\leq 5\%$] (Table
219 S7) and many of these [127/200] are uncharacterized in CF (Table S8).

220

221 **More than one CF-causing variants detected in participants**

222

223 With CF being an autosomal recessive disease, we explored the prevalence of participants with
224 two CF-causing variants and those heterozygous for a CF-causing variant and a VCC or high
225 impact variant. We found 16 UKBB participants possessing at least two CF-causing variants
226 (Table 3). These participants were found exclusively in the CSA [n=1], EUR (n=9), and MIX
227 [n=6] populations. The CSA participant was homozygous for V456A, but various CF-causing
228 variant combinations were found in the EUR and MIX populations. F508del was the most
229 common CF-causing INDEL detected across the populations. Interestingly, the two individuals

230 homozygous for F508del were also diagnosed as having CF in UKBB. Many of the participants
231 with two CF-causing *CFTR* variants had combinations associated with pancreatic sufficiency
232 (96%). The three variant combinations associated with pancreatic insufficiency were
233 R851X/F508del and F508del/F508del [n=2] (Table 3). We found 155 participants that harbored
234 a CF-causing variant and a VCC (Table 4). These participants were found only in the EUR
235 [n=145] and MIX [n=10] populations. R117H was the most reported VCC [49/154] variant in
236 this subset of participants, and its combination with F508del was the most common found in the
237 EUR [38/145] and MIX [5/10] populations. L967S also featured as a common VCC variant. Its
238 combination with F508del was exclusively reported in the EUR population and dominated as the
239 second most common combination (Table 4). Since high impact variants are potentially
240 functionally relevant, we assessed their distribution across the ancestries. Among the 200 high
241 impact variants detected in the participants, the highest number was found in EUR population
242 [150/200] while the least number of these variants [4/200] were reported in both the AMR and
243 EAS populations [Table S7]. We also found a high number of participants heterozygous for a
244 CF-causing and high impact variant [Table S9]. These findings warranted us to interrogate the
245 association of these variant with participants health outcomes.

246

247 **Classical CF phenotypes reported in participants with two CF-causing variants and VCCs**

248

249 Phenotype analysis performed using hospital in-patient records indicated that most of the
250 participants with at least two CF-causing variants presented classical CF phenotypes. Eight
251 (International Classification of Diseases 10th Revision (ICD-10) codes were significantly
252 [FDR<0.05] enriched in participants with two CF-causing variants (Fig. 5A). Specifically, ICD
253 codes directly associated with CF [E840, 848, and 849] and colonization with *Pseudomonas*
254 *aeruginosa* [B965] were among the top significantly enriched terms. Interestingly, ICD Code
255 U837 which corresponds to Resistance to Multiple Antibiotics also featured among the top 10
256 enriched phenotypes. Although bronchiectasis, a classical consequence of dysfunctional *CFTR*,
257 was listed among the top, it did not meet our significance threshold when corrected for multiple
258 testing (Fig. 4). With VCC variants resulting in variable clinical outcomes when combined with
259 another CF-causing variant but not often diagnosed as CF, we assessed the phenotypes of
260 participants that fall within this category. Our analysis indicated that six ICD codes were
261 significantly [FDR<0.05] associated with CF [E840 and 848]; of these, pulmonary phenotypes
262 [J47, B441, J998, and J440] were significantly enriched in these participants (Fig. 5B). In an
263 attempt to characterize the clinical relevance of high impact variants, we then explored the
264 phenotypes of participants with one CF-causing variant and a high impact variant not known to
265 cause CF (n=272). Surprisingly, the only significant association [FDR<0.05] returned was
266 Melanoma in situ of the trunk which is not a classical CF phenotype. Aspergillosis was featured
267 among the top 10 most enriched phenotypes, but this association failed to meet our significance
268 threshold (Table S9). Further analysis deployed to compare differences of spirometry measures
269 (FEV1 and FVC) between participant with two CF-causing variants and the general population
270 returned no significant results.

271

272 **Chronological age of CF-causing variants**

273

274 GEVA reported the estimated age of over 4,000 variants within the *CFTR* gene locus (Table
275 S10). Only four of the dated variants were CF-causing [rs139729994 (3600G>A), rs193922500

276 (V456A), rs76713772 (1717-1G>A), rs77646904 (V520F)] (Fig. S1). The youngest variant
277 among these, 1717-1G->A, appears to have originated ~425 generations ago (~10,600 years)
278 while the earliest variant, V456A, originated about ~215 generations ago (~5,300 years).
279 Interestingly, V520F and 3600G>A which appeared ~338 and 270 generations ago featured
280 among the most common CF-causing variants reported in EAS population (Fig. 3). When the
281 allele frequencies of the CF-causing variants were compared with variants of similar ages within
282 the respective populations in AVA, there were no significant differences. Thus, an indication that
283 the variants were not under any selective pressure at the time of origination.

284

285 Discussion

286

287 Molecular diagnosis of CF is confounded by sparse knowledge about the prevalence of
288 pathogenic and non-pathogenic *CFTR* variants across ancestries, including those that are
289 potentially ancestry specific. In this study, we interrogated whole-exome sequence datasets
290 generated from over 450, 000 UKBB participants to characterize the distribution of *CFTR*
291 variants across six ancestries [AFR, CSA, EAS, EUR, AMR, and MID], including an
292 uncharacterized group (MIX). We report, for the first time, the detection of over 4000 *CFTR*
293 variants across the ancestries. Among the detected ancestry-specific variants were several
294 variants of clinical and functional relevance. Phenotypic characterization of participants
295 harboring multiple combinations of these variants reported indications associated with CF and its
296 classical pulmonary phenotypes.

297

298 Although previous efforts driven by the CFTR1/2 team have led to the identification of just over
299 2000 *CFTR* variants within the CF population¹⁰, our report of over 4000 *CFTR* variants is a
300 small fraction of potential variants that could exist within the *CFTR* gene. *In silico* analyses of
301 external variant browsers such as BRAVO¹⁷, which is based on ~132k WGS generated by the
302 TOPMed consortium, and All of Us Research Hub²⁵, which is based on ~98k WGS and ~165k
303 genotyping arrays reports a total of 56,448 and 39,797 *CFTR* variants, respectively. The higher
304 number of *CFTR* variants captured in these databases are perhaps due to WGS having the
305 capability of capturing the entire length of the *CFTR* gene while WES captures mostly exonic
306 regions. Overall, the vast majority of the *CFTR* variants found in UKBB [52%], All of Us
307 Research Hub [89%], and BRAVO [96%], were intronic variants. Overall, c.1408G>A (V470M
308 [rs213950]) was the most common allele (Table S1). Although in GRCH37/38 the reference
309 allele is a G, the variation associated with increased risk of chronic pancreatitis indicate at this
310 position is c.1408A>G (M470V)²⁶. Regardless, a combination of this variant with another CF-
311 causing does not cause CF (<https://cftr2.org>). CF genetic diagnosis is largely based on
312 characterized *CFTR* variants, most of which are reported in Europeans^{4,10}. Although the UKBB
313 is dominated by people of European ancestry, which could potentially influence estimates of
314 genetic variations²⁷, CF is one of the most common life-threatening genetic disease reported in
315 this population²⁸. In line, with previous reports we also found the highest number of *CFTR*
316 variants in EUR population. Thus, the mutation rate of the *CFTR* gene is likely higher in the
317 EUR than other populations. The aspect of precision medicine for *CFTR* modulators is bound to
318 make higher impact in this CF population with such variable numbers of mutations.

319

320 Many clinically (CF-causing and VCC) and functionally (High-Impact) relevant variants were
321 identified following cross-ancestry annotation of all detected variants. As their global

322 distribution and frequencies are poorly understood, we characterized the most common and
323 unique CF-causing variants across populations. Indeed, like previous reports^{13,29,30}, F508del was
324 the most common CF-causing variant reported globally. While F508del represented ~90% of the
325 CF-causing variants found in EUR, our findings indicate its prevalence is much lower in other
326 populations (Fig. 3). Thus, F508del may be the major cause of CF in EUR population, but not for
327 people of all ancestries. Lower frequencies of F508del in non-European populations have been
328 reported in several studies, and the spectrum of variants causing CF in such populations
329 varies^{31,32}. For example, G970D has been reported as the most common cause of CF in the
330 Chinese population¹². This variant, along with V520F and 3600G>A, featured among the top
331 three common variants we found in the EAS group and were more common than F508del (Fig.
332 3). Although we found V520F to be the most common in EAS, it is conceivable G970D is more
333 common to China than V520F since the EAS group examined in this study expands beyond the
334 Chinese population. Like other non-EUR populations, CF is less frequent or likely underreported
335 in Africans¹⁰. Although F508del was the most common CF-causing variant found in the AFR
336 group, 3120+1G>A which is predominantly found in Africans¹⁰, was also found to be common
337 to the AFR group (Fig. 3). Also, in South Asians, CF is thought to be less frequent. Although
338 F508del represents about 40-50% of cases, V456A the second most common CF-causing variant
339 we found in CSA (Fig.3), is a well-characterized cause of CF in South Asians^{33,34}. Some other
340 common variants found in non-EUR populations, such as AFR (3791delC), CSA (1525-1G>A),
341 and EAS, were ancestry-specific. By comparing our findings with BRAVO reports¹⁷, we found
342 these variants were also rare and exclusively possessed by similar populations. Unfortunately, we
343 found most of the common CF-causing variants found in non-European populations are not
344 included in the American College of Medical Genetics list of 23 variants recommended for CF
345 carrier screening³⁵.

346
347 Numerous hypotheses have been put forward to explain the persistence of CF-causing variants
348 across populations, despite their detrimental impact in life outcomes. A leading hypothesis
349 postulates the heterozygous advantage of CF-causing variants against infectious diseases, such as
350 cholera, typhoid fever, or tuberculosis³⁶. For example, the most common CF-causing variant,
351 rs113993960 (F508del), has been estimated to have arisen approximately 600 generations
352 ago^{37,38} and its rapid increase in Europeans (MAF = 1.6%) was linked to the tuberculosis
353 pandemic of the seventeenth century, against which it is thought to increase resistance in
354 mutation carriers³⁶. Among other variants with estimated age data in AVA, we did not detect any
355 variant that has rapidly increased in allele frequency as the F508del variant in Europeans.
356 Although, V520F and 3600G>A are more common CF-causing variants in EAS population than
357 F508del (Fig. 3), their allele frequencies were not significantly different from other variants of
358 similar ages in EAS population. Thus, F508del is likely a unique variant that confers selective
359 pressure, which has kept its frequency relatively high.

360
361 Beyond CF-causing variants, we also found several VCCs, and High Impact variants not known
362 to cause CF but predicted to severely impact protein function. R117H the common VCC, which
363 results in reduced single-channel activity and open probability³⁹, was the most prevalent VCC
364 detected across the ancestries (Table S5). Clinical manifestations of patients with R117H are
365 heterogenous and largely influenced by its combination with another CF-causing variant or other
366 common variants of poly-T tract (5T, 7T, and 9T). R117H-T5 is reported to potentially result in
367 less functional CFTR and pancreatic sufficiency when patients are homozygous for this

368 combination or when found in compound heterozygosity with a CF-causing variant⁴⁰. Most of
369 the predicted High Impact variants found in our study were already known to cause CF (deduced
370 with the CFTR2 annotation). After excluding these variants, we still found some uncharacterized
371 High Impact variants. Taken together, these findings indicated a handful of UKBB participants
372 harbored clinically and functionally relevant *CFTR* variants.

373
374 Since CF is an autosomal recessive disease, we characterized UKBB participants with more than
375 one clinical or functional relevant variant. Interestingly, we found sixteen UKBB participants
376 with at least two CF-causing variants. Collectively, these participants reported ICD codes
377 associated with classical CF phenotypes [Fig. 5A], despite the fact only two participants
378 (F508del homozygotes) were diagnosed as CF in UKBB electronic health records. Undiagnosed
379 cases of CF, especially in patients with milder forms of the disease caused by rare pathogenic
380 genotypes, is not uncommon⁴¹. Most of the variant combinations reported in our findings likely
381 result in pancreatic sufficiency, except for one participant harboring the F508del/R851X
382 genotype that results in pancreatic insufficiency. Although the F508del/R851X participant was
383 not diagnosed as CF in UKBB, the retrieved ICD10 code indicated the participant suffered
384 majorly from Chest Pain [R074]. This variant combination, which was detected in an EUR
385 participant, is less common in CF, with a frequency of only ~0.0002 recorded in CFTR2
386 database [<https://cftr2.org>]. Another rarer genotype with ~0.00002 frequency in CFTR2 which
387 we found was V456A/V456A. This variant is known to cause CF in South Asians³³ and likewise
388 was detected in a CSA participant. Some of the detected pathogenic genotypes, such as
389 L206W/S945L and R1066H/1154insTC, have not yet been recorded in CFTR2. Taken together,
390 a possible explanation for the undiagnosed cases captured in our study may be due to the
391 rareness of the variants and milder presentation of the disease.

392
393 We recorded about 10-fold more participants harboring a CF-causing variant and a VCC, than
394 those with two CF-causing variants (154 vs 16, respectively). VCCs contribute majorly to the
395 complexities associated with genetic diagnosis of CF, because clinical outcomes vary widely
396 across patient populations⁴². For this sub-group of participants, we also found significant
397 enrichment of ICD codes associated with classical CF (ICD Codes: E840 and E848) and
398 pulmonary phenotypes, including bronchiectasis, Aspergillosis, and COPD with acute lower
399 respiratory infections. Participants with these genotypes were reported only in EUR and MIX
400 populations, and the most dominant combination was F508del/R117H. This variant combination
401 on its own does not cause CF but when in *cis* with 5T ploy-Tract variant, then CF diagnosis is
402 likely⁴⁰. About 1,310 patients have the F508del/R117H variant combination in CFTR2 and 22%
403 of these are pancreatic insufficient. Meanwhile, 80 patients have the F508del/R117H;5T
404 combination and 33% of them are pancreatic insufficient [<https://cftr2.org>]. Although these
405 participants are not diagnosed as CF, their inclusion as patients in CFTR2 suggest they suffer the
406 burden of dysfunctional CFTR. The F508del/L967S genotype, which was the second most
407 common found in this subgroup (CF-causing+VCC), was rare in CFTR2. Only seven patients
408 had this combo and clinical outcome, though variable, is largely expected to be pancreatic
409 sufficient. The identification of several UKBB participants harboring such variant combinations
410 and showing CF-like symptoms, is therefore an indication that a number of undiagnosed or
411 unreported CF cases may benefit from modulator therapies. Additionally, we found no
412 differences in pulmonary measures (FEV and FVC) between participants with two CF-causing
413 variants and the general population. Given that the UKBB recruited individuals that were over 40

414 years old, the individuals with multiple clinically relevant *CFTR* genotypes likely have a milder
415 form of the disease with a different disease etiology that nevertheless could be potentially either
416 treated or prevented by appropriate *CFTR*-based interventions.

417
418 A major drawback to existing knowledge about pathogenic *CFTR* variants is that most studies
419 have been conducted in European populations. However, CF is also prevalent in people from
420 other ancestries, who may harbor uncharacterized unique variants⁴. Even in individuals of
421 European ancestry, where most of the studies have been conducted, there are several *CFTR*
422 variants with unknown clinical significance. In attempt to decipher the clinical relevance of
423 uncharacterized high impact *CFTR* variants detected across the UKBB population, we performed
424 a phenotype analysis of participants with a CF-causing variant and a non-CF-causing
425 (uncharacterized) high impact variant but found no interesting association. Databases such as the
426 UKBB offer an unprecedented opportunity to correlate uncharacterized genotype and phenotype
427 information at a population scale. Although we utilized the UKBB resource to characterize the
428 ancestral diversity of *CFTR* variants, most of the participants are of European ancestry. Recent
429 genome sequencing efforts that included more diverse populations, such as the NIH All of Us
430 and ToPMed programs, report even more *CFTR* variants in comparison to UKBB. However,
431 limited access to phenotype information at an individual level makes it challenging to unravel the
432 clinical relevance of the detected *CFTR* variants.

433
434 Here, we aimed to characterize the natural variation that exists within the *CFTR* gene that is not
435 limited to persons with CF. The large majority (>95%) of variants surveyed have allele
436 frequencies less than 1% and ~9% of all variations are indels. While sequencing technology has
437 improved in accuracy and quality control filters have removed many of the problematic variants,
438 identification, and interpretation of novel rare variants (specifically singletons and doubletons)
439 would warrant more precise methods (such as targeted sequencing) for confirmation. Another
440 limitation of our study is the lack of precise CF-related phenotypes, such as FEV1% predicted
441 and sweat chloride levels in the electronic health records extracted from UKBB. Although we
442 captured some spirometry measures, we lacked information relevant to deduce FEV1% scores
443 and there were no records of sweat chloride levels for the participants.

444
445 In summary, we have leveraged the UKBB resource to comprehensively characterize the broad
446 spectrum of *CFTR* variants across ancestries. Until now, most efforts aimed at identifying and
447 characterizing *CFTR* variants have focused more on surveying CF populations in a relatively
448 smaller sample size. For the first time, we report the detection of over 4000 *CFTR* variants,
449 which nearly doubles the number of variants reported by CFTR1/2. However, the higher number
450 of variants reported by the NIH All of Us and ToPMed diverse sequencing programs suggests
451 the 4000 *CFTR* variants represents only a small fraction of variants that exist in the *CFTR* gene.
452 The identification of several ancestry-specific variants, including uncharacterized functionally
453 relevant variants, warrants the need for CF screening panels to consider ethnic specificities. The
454 presentation of classical CF phenotypes seen in some non-CF diagnosed participants, with CF-
455 causing and VCC variants, indicates they may benefit from current *CFTR* therapies. As more
456 diverse *CFTR* sequences are becoming increasingly available in genomic databases, future
457 studies are encouraged to leverage such resources to characterize the clinical relevance of high
458 impact variants not previously known to cause CF.

459

460 **Supplementary information**

461
462 Supplementary Tables 1-7: **Table S1.** Detected *CFTR* variants across UKBB populations; **Table**
463 **S2.** Unique CF-causing variants; **Table S3.** CF-causing *CFTR* variants; **Table S4.** Variable
464 clinical consequence variants detected across UKBB ancestries; **Table S5.** Unique CF-causing
465 variants; **Table S6.** Common varying clinical consequences (VCC) variants; **Table S7.** *CFTR*
466 variants annotated as having High Impact; **Table S8.** Uncharacterized high impact variants;
467 **Table S9.** High impact and CF-Causing variants. **Table S10.** Dated variants within the *CFTR*
468 gene locus
469 Supplementary Figures: **Figure S1.** Venn diagram showing genealogical estimation of variant
470 age (GEVA) analysis of CF-causing variants

471 472 **Data availability**

473
474 Access to UK biobank resource is available by application (<http://www.ukbiobank.ac.uk/>). The
475 UK biobank exome sequences analyzed in this study were obtained under application 26401.

476 477 **Ethics declarations**

478
479 Competing interests: The authors declare no competing interest. The design, study conduct, and
480 financial support for this research were provided by AbbVie. AbbVie participated in the
481 interpretation of data, review, and approval of the publication. JEI, ML, BRG, SRP, FR, PK, RT,
482 PD, AS, JFW, and AV are employees of AbbVie. HL is an employee of University of Wisconsin
483 School of Medicine and Public Health and provided her expertise on CF genetics and
484 translational research (no funding to disclose).

485 486 **Acknowledgement**

487
488 We would like to acknowledge and extend our gratitude to the AbbVie GRC Internal Reviewers;
489 Anneke Den Hollander and Nizar Smaoui.

490 491 **References**

- 492
493 1. Raraigh, K.S. *et al.* Complete *CFTR* gene sequencing in 5,058 individuals with cystic
494 fibrosis informs variant-specific treatment. *J Cyst Fibros* **21**, 463-470 (2022).
495 2. Desportes, K.A. & Donaldson, S.H. Current state of *CFTR* modulators for treatment of
496 Cystic Fibrosis. *Curr Opin Pharmacol* **65**, 102239 (2022).
497 3. Griese, M. *et al.* Safety and Efficacy of Elexacaftor/Tezacaftor/Ivacaftor for 24 Weeks or
498 Longer in People with Cystic Fibrosis and One or More F508del Alleles: Interim Results
499 of an Open-Label Phase 3 Clinical Trial. *Am J Respir Crit Care Med* **203**, 381-385
500 (2021).

- 501 4. Guo, J., Garratt, A. & Hill, A. Worldwide rates of diagnosis and effective treatment for
502 cystic fibrosis. *J Cyst Fibros* **21**, 456-462 (2022).
- 503 5. Corvol, H. *et al.* Genome-wide association meta-analysis identifies five modifier loci of
504 lung disease severity in cystic fibrosis. *Nat Commun* **6**, 8382 (2015).
- 505 6. Petrova, N. *et al.* Ethnic Differences in the Frequency of CFTR Gene Mutations in
506 Populations of the European and North Caucasian Part of the Russian Federation. *Front*
507 *Genet* **12**, 678374 (2021).
- 508 7. Schrijver, I. *et al.* The Spectrum of CFTR Variants in Nonwhite Cystic Fibrosis Patients:
509 Implications for Molecular Diagnostic Testing. *J Mol Diagn* **18**, 39-50 (2016).
- 510 8. Palomaki, G.E., FitzSimmons, S.C. & Haddow, J.E. Clinical sensitivity of prenatal
511 screening for cystic fibrosis via CFTR carrier testing in a United States panethnic
512 population. *Genet Med* **6**, 405-14 (2004).
- 513 9. Shteinberg, M., Haq, I.J., Polineni, D. & Davies, J.C. Cystic fibrosis. *Lancet* **397**, 2195-
514 2211 (2021).
- 515 10. Stewart, C. & Pepper, M.S. Cystic Fibrosis in the African Diaspora. *Ann Am Thorac Soc*
516 **14**, 1-7 (2017).
- 517 11. Quint, A., Lerer, I., Sagi, M. & Abeliovich, D. Mutation spectrum in Jewish cystic
518 fibrosis patients in Israel: implication to carrier screening. *Am J Med Genet A* **136**, 246-8
519 (2005).
- 520 12. Tian, X. *et al.* p.G970D is the most frequent CFTR mutation in Chinese patients with
521 cystic fibrosis. *Human Genome Variation* **3**, 15063 (2016).
- 522 13. Ni, Q. *et al.* Systematic estimation of cystic fibrosis prevalence in Chinese and genetic
523 spectrum comparison to Caucasians. *Orphanet J Rare Dis* **17**, 129 (2022).
- 524 14. Levy, H. *et al.* Identification of molecular signatures of cystic fibrosis disease status with
525 plasma-based functional genomics. *Physiol Genomics* **51**, 27-41 (2019).
- 526 15. Grosse, S.D. *et al.* Newborn screening for cystic fibrosis: evaluation of benefits and risks
527 and recommendations for state newborn screening programs. *MMWR Recomm Rep* **53**, 1-
528 36 (2004).
- 529 16. Coverstone, A.M. & Ferkol, T.W. Early Diagnosis and Intervention in Cystic Fibrosis:
530 Imagining the Unimaginable. *Front Pediatr* **8**, 608821 (2020).
- 531 17. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed
532 Program. *Nature* **590**, 290-299 (2021).
- 533 18. Moisan, S. *et al.* Analysis of long-range interactions in primary human cells identifies
534 cooperative CFTR regulatory elements. *Nucleic Acids Research* **44**, 2564-2576 (2015).
- 535 19. Backman, J.D. *et al.* Exome sequencing and analysis of 454,787 UK Biobank
536 participants. *Nature* **599**, 628-634 (2021).
- 537 20. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122 (2016).
- 538 21. Cingolani, P. *et al.* A program for annotating and predicting the effects of single
539 nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster*
540 strain w1118; iso-2; iso-3. *Fly (Austin)* **6**, 80-92 (2012).
- 541 22. Ideozu, J.E. *et al.* Increased Expression of Plasma-Induced ABCC1 mRNA in Cystic
542 Fibrosis. *Int J Mol Sci* **18**(2017).
- 543 23. McKay, I.R. & Ooi, C.Y. The Exocrine Pancreas in Cystic Fibrosis in the Era of CFTR
544 Modulation: A Mini Review. *Front Pediatr* **10**, 914790 (2022).
- 545 24. Albers, P.K. & McVean, G. Dating genomic variants and shared ancestry in population-
546 scale sequencing data. *PLoS Biol* **18**, e3000586 (2020).

- 547 25. Denny, J.C. *et al.* The "All of Us" Research Program. *N Engl J Med* **381**, 668-676 (2019).
- 548 26. Zhou, D., Bai, R. & Wang, L. The Cystic Fibrosis Transmembrane Conductance
549 Regulator 470 Met Allele Is Associated with an Increased Risk of Chronic Pancreatitis in
550 Both Asian and Caucasian Populations: A Meta-Analysis. *Genetic Testing and Molecular*
551 *Biomarkers* **24**, 24-32 (2020).
- 552 27. Sun, Q. *et al.* Analyses of biomarker traits in diverse UK biobank participants identify
553 associations missed by European-centric analysis strategies. *J Hum Genet* **67**, 87-93
554 (2022).
- 555 28. Blanchard, A.C. & Waters, V.J. Opportunistic pathogens in cystic fibrosis: epidemiology
556 and pathogenesis of lung infection. *Journal of the Pediatric Infectious Diseases Society*
557 **11**, S3-S12 (2022).
- 558 29. Lima, E.d.S., Pezzin, L.S., Fensterseifer, A.C. & Pinto, L.A. Frequency of CFTR variants
559 in southern Brazil and indication for modulators therapy in patients with cystic fibrosis.
560 *Genetics and Molecular Biology* **45**(2021).
- 561 30. Erdoğan, M. *et al.* The Genetic Analysis of Cystic Fibrosis Patients with Seven Novel
562 Mutations in the CFTR Gene in the Central Anatolian Region of Turkey. *Balkan Medical*
563 *Journal* **38**, 357 (2021).
- 564 31. Ortiz, S.C. *et al.* Spectrum of CFTR gene mutations in Ecuadorian cystic fibrosis
565 patients: the second report of the p.H609R mutation. *Mol Genet Genomic Med* **5**, 751-757
566 (2017).
- 567 32. Siryani, I. *et al.* Distribution of Cystic Fibrosis Transmembrane Conductance Regulator
568 (CFTR) Mutations in a Cohort of Patients Residing in Palestine. *PLoS One* **10**, e0133890
569 (2015).
- 570 33. Uppaluri, L., England, S.J. & Scanlin, T.F. Clinical evidence that V456A is a Cystic
571 Fibrosis causing mutation in South Asians. *J Cyst Fibros* **11**, 312-5 (2012).
- 572 34. Indika, N.L.R. *et al.* Phenotypic spectrum and genetic heterogeneity of cystic fibrosis in
573 Sri Lanka. *BMC Medical Genetics* **20**, 89 (2019).
- 574 35. Deignan, J.L. *et al.* CFTR variant testing: a technical standard of the American College of
575 Medical Genetics and Genomics (ACMG). *Genetics in Medicine* **22**, 1288-1295 (2020).
- 576 36. Poolman, E.M. & Galvani, A.P. Evaluating candidate agents of selective pressure for
577 cystic fibrosis. *J R Soc Interface* **4**, 91-8 (2007).
- 578 37. Morris, A.P., Whittaker, J.C. & Balding, D.J. Fine-scale mapping of disease loci via
579 shattered coalescent modeling of genealogies. *Am J Hum Genet* **70**, 686-707 (2002).
- 580 38. Slatkin, M. & Bertorelle, G. The use of intraallelic variability for testing neutrality and
581 estimating population growth rate. *Genetics* **158**, 865-74 (2001).
- 582 39. Yu, Y.C., Sohma, Y. & Hwang, T.C. On the mechanism of gating defects caused by the
583 R117H mutation in cystic fibrosis transmembrane conductance regulator. *J Physiol* **594**,
584 3227-44 (2016).
- 585 40. Castellani, C. *et al.* Consensus on the use and interpretation of cystic fibrosis mutation
586 analysis in clinical practice. *J Cyst Fibros* **7**, 179-96 (2008).
- 587 41. Sagesse, G.J., Yadava, S. & Mandava, A. Atypical Cystic Fibrosis: Diagnosis at the Age
588 of 57 Years. *Cureus* **12**, e10863 (2020).
- 589 42. De Wachter, E., Thomas, M., Wanyama, S.S., Seneca, S. & Malfroot, A. What can the
590 CF registry tell us about rare CFTR-mutations? A Belgian study. *Orphanet Journal of*
591 *Rare Diseases* **12**, 142 (2017).
- 592

593
594

Approximate global ancestry	Sample Size, n (%)	Gender: Male, n (%)	Age in Years, Median (IQR)	CFTR Variants, n (%)	Unique CFTR Variants, n (%)
African (AFR)	6278 (1.4%)	2582 (41%)	62 (57,70)	417 (9.9%)	138 (33.1%)
Central South Asia (CSA)	8470 (1.9%)	4572 (54%)	65 (58, 72)	471 (11.2%)	177 (37.6%)
East Asia (EAS)	2608 (0.6%)	886 (34%)	63 (57,70)	266 (6.3%)	104 (39.1%)
European (EUR)	397,073 (87.3%)	182164 (46%)	70 (62,75)	3192 (76.1%)	2212 (69.3%)
American/Admixed American (AMR)	950 (0.2%)	334 (35%)	63 (57,71)	151 (3.6%)	23 (15.2%)
Middle East (MID)	1498 (0.3%)	872 (58%)	63 (57,70)	222 (5.3%)	46 (20.7%)
MIX	37850 (8.3%)	16450 (43%)	67 (60,67)	1333 (31.8%)	381 (28.6%)

595
596
597
598
599
600
601
602
603
604
605
606

Table 1. Diversity of CFTR variants across ancestries

Table 2. Number of participants with CF-causing variants across ancestries

Ancestries	Sample Size in UKBB	CFTR Variants	CF-Causing	Varying Clinical Consequences (VCC)	High Impact Variants
AFR	6278	417	22	12	16
AMR	950	151	7	10	4
CSA	8470	471	17	6	15
EAS	2608	266	8	6	4
EUR	397,073	3,192	116	29	150
MID	1498	222	8	12	7
MIX	37850	1,333	85	29	70

607
608
609

Table 3. Participants with two CF-causing variants

Ancestry	Variant1	Variant2	Number of Variant	Predicted Phenotype	CFTR2 Participants* [n]
EUR	L206W	R553X	1	Pancreatic Sufficient	3
	L206W	S945L	1	Pancreatic Sufficient	0
	R851X	F508del	1	Pancreatic Insufficient	16
	R1066H	F508del	1	Pancreatic Sufficient	56
	R1066H	1154insTC	1	Pancreatic Sufficient	0
	P67L	G551D	1	Pancreatic Sufficient	12

	P67L	F508del	1	Pancreatic Sufficient	186
	711+3A->G	F508del	1	Pancreatic Sufficient	42
	F508del	F508del	1	Pancreatic Insufficient	33984
MIX	F508del	D110H	2	Pancreatic Sufficient	34
	F508del	L206W	1	Pancreatic Sufficient	235
	F508del	F508del	1	Pancreatic Insufficient	33984
	3600G->A	F508del	1	Pancreatic Sufficient	15
	E474K	F508del	1	Pancreatic Sufficient	3
CSA	V456A	V456A	1	Pancreatic Sufficient	2

610 **Table 4. UKBB participants with a CF-causing and variable clinical consequence (VCC) variant**
611

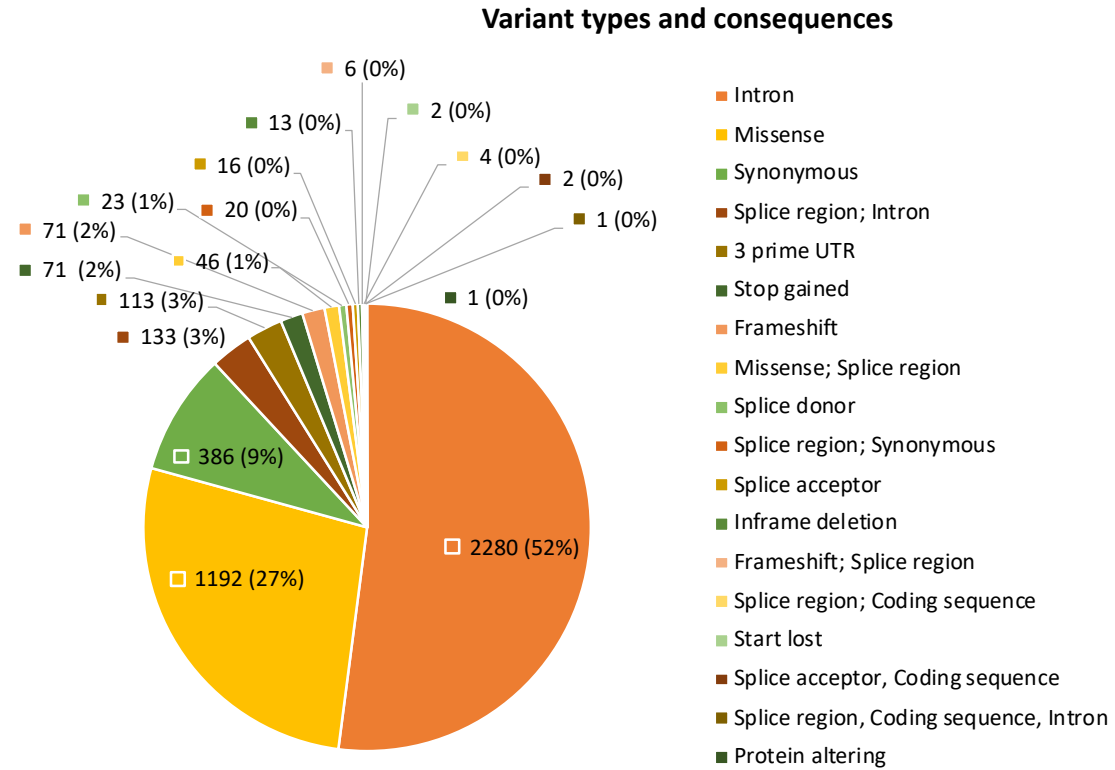
Ancestry	Variant1	Variant2	Count
EUR	R117H	F508del	38
	L967S	F508del	30
	D443Y	F508del	13
	P750L	F508del	12
	621+3A->G	F508del	5
	R334Q	F508del	4
	P750L	G551D	3
	R258G	F508del	3
	D1152H	F508del	2
	F1052V	F508del	2
	L967S	1717-1G->A	2
	L967S	G551D	2
	L967S	W1282X	2
	Q1291H	F508del	2
	R1070W	F508del	2
	D1270N	F508del	1
	D443Y	1898+1G->A	1
	D443Y	N1303K	1
	D443Y	Q493X	1
	D443Y	R553X	1
	F575Y	1154insTC	1
	L967S	1138insG	1
	L967S	1898+1G->A	1
	L967S	3659delC	1
	M265R	F508del	1
	P750L	1461ins4	1
	P750L	R560T	1
	P750L	W846X	1
	R1070Q	F508del	1
	R1070W	G542X	1
	R117G	F508del	1
	R117H	1898+1G->A	1
	R117H	2184delA	1
	R117H	3272-26A->G	1
R117H	I507del	1	
R117H	R347H	1	

	R117H	R347P	1
	R352W	N1303K	1
MIX	R117H	F508del	5
	621+3A->G	F508del	1
	D443Y	F508del	1
	D1152H	1898+1G->A	1
	F1052V	F508del	1
	Q1476X	N1303K	1

612

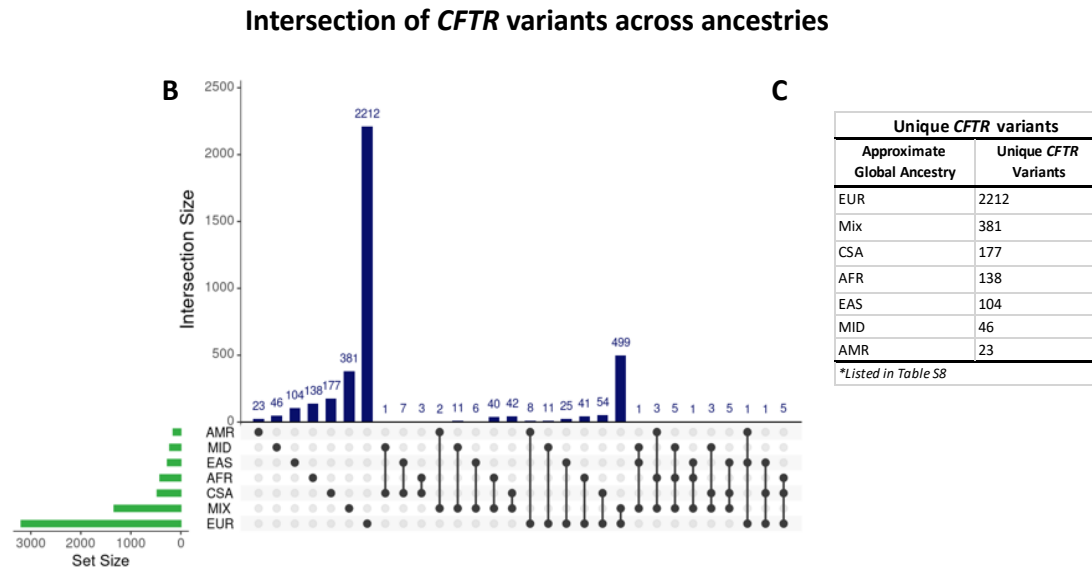
613

Fig. 1: Most *CFTR* variants are intronic



Several variant types with varying consequences were identified in the *CFTR* gene. The labels in the chart represent the number and percentage of the Intronic variants (highlighted in orange) clearly dominated over half of the captured variants despite Whole Exome Sequencing technology capturing mostly exomes.

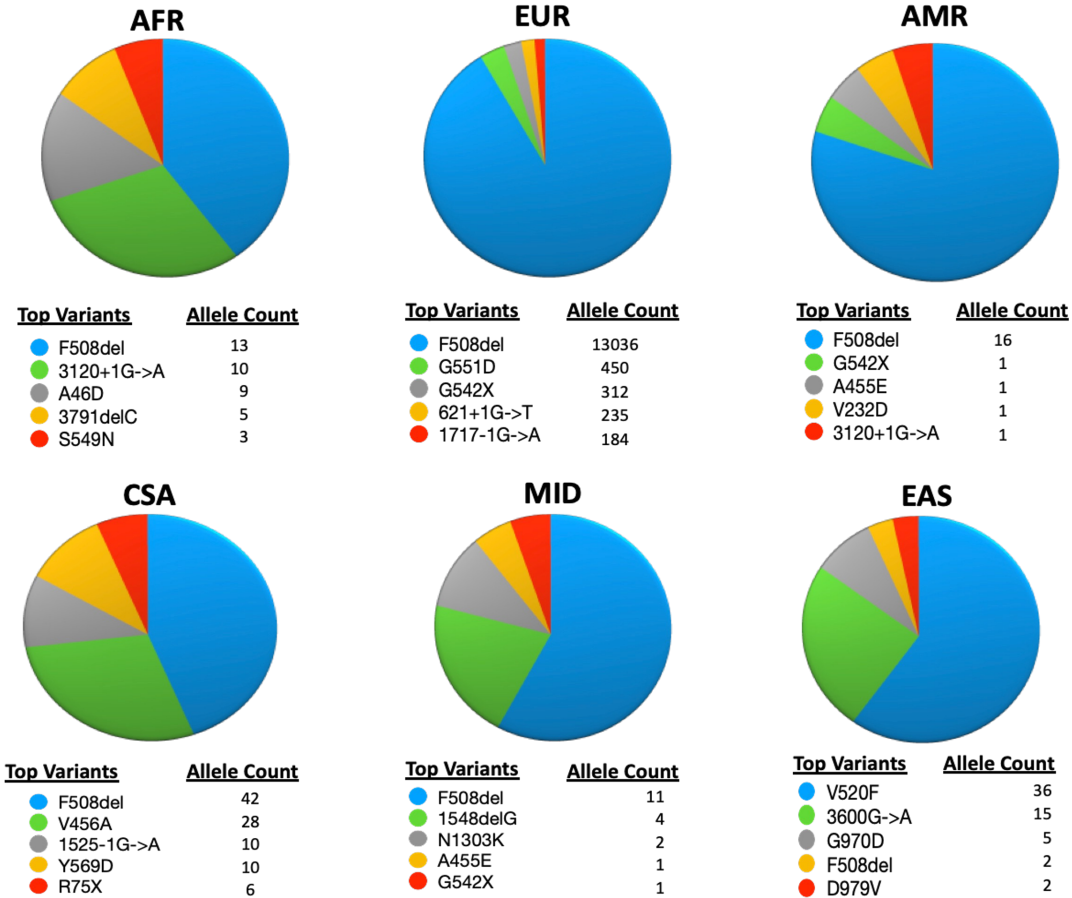
Fig. 2: Ancestry-specific variants characterized



A) Graph of total number of variants (x-axis; Set Size) detected in each ancestry (y-axis). **B)** Intersection of variants detected across the ancestries. Each column corresponds to the number of variants in each intersection. Ancestries present in each intersection are represented by the black dots. Intersections with single dots represents ancestry-specific variants. **C)** A list of ancestry-specific variants. The highest number of unique *CFTR* variants (2212/3192) was detected in Europeans.

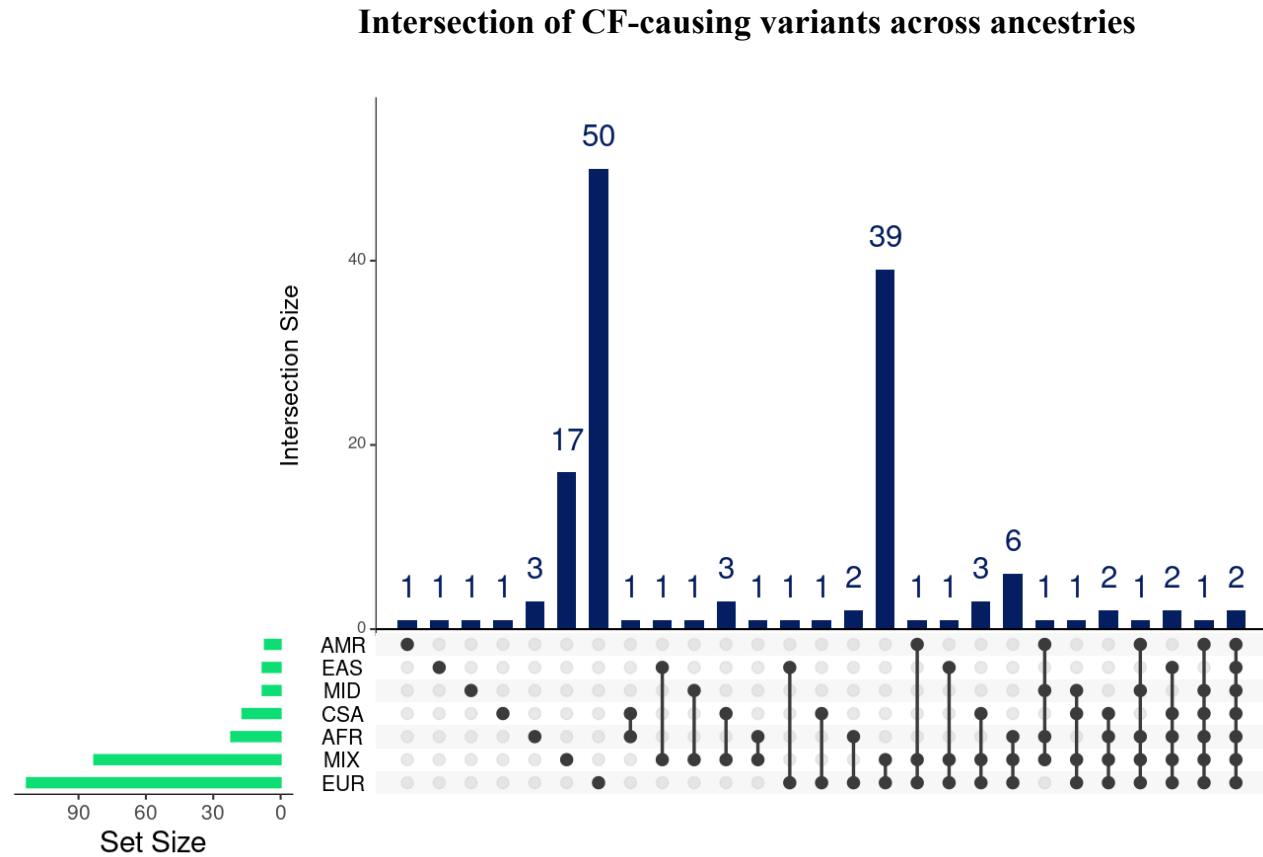
Fig. 3: Common CF-causing variants in each ancestry represented in pie charts

Common CF-causing variants across ancestries



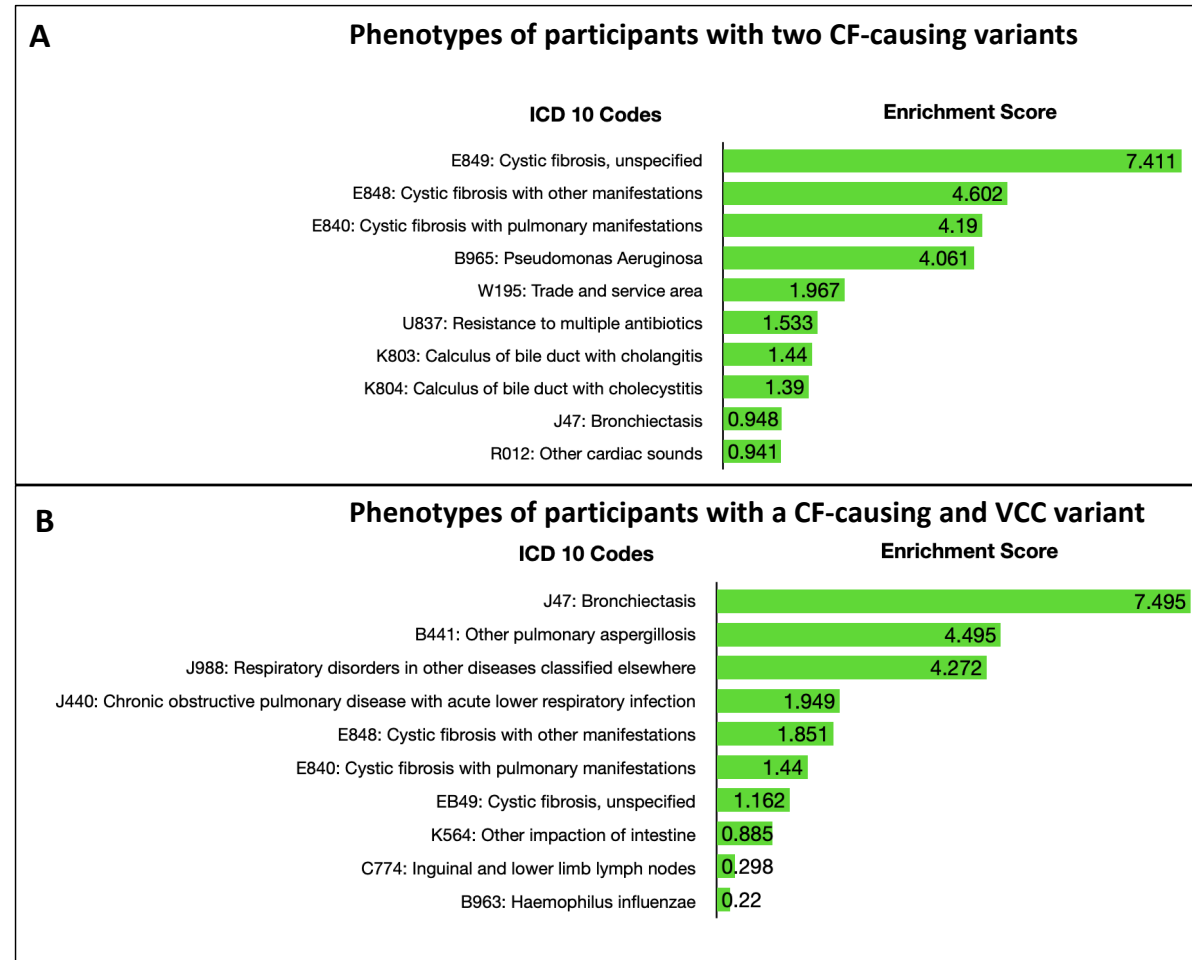
For each ancestry, slices vary by color, with blue representing the most abundant CF-causing variant. F508del was the most common CF-causing variant detected in all ancestries, except for East Asia where V520F was the most common.

Fig. 4: Ancestry-specific CF-causing variants characterized



A) Graph of total number of variants (x-axis; Set Size) detected in each ancestry (y-axis). **B)** Intersection of variants detected across the ancestries. Each column corresponds to the number of variants in each intersection. Ancestries present in each intersection are represented by the black dots. Intersections with single dots represents ancestry-specific CF-causing variants. **C)** A list of ancestry-specific CF-causing variants. The highest number of unique CF-causing variants (n=50) was detected in Europeans.

Fig. 5: Classical CF phenotypes enriched in participants with >1 clinically relevant *CFTR* variant



A) Enriched phenotypes observed in patients with two CF-causing variants **B)** Enriched phenotypes observed in patients with one CF-causing and a high impact variant. The top 10 enriched phenotypes are represented in bar graphs. The negative logarithm of the Bonferroni adjusted p-value was used to deduce the enrichment scores shown in the bars. Enrichment score >1.3 corresponds to a significance threshold of $p < 0.05$.