Systematic estimation of cystic fibrosis prevalence in Chinese and genetic spectrum comparison to Caucasians

Qi Ni

Children's Hospital of Fudan University

Xiang Chen

Children's Hospital of Fudan University

Ping Zhang

Children's Hospital of Fudan University

Lin Yang

Children's Hospital of Fudan University

Yulan Lu

Children's Hospital of Fudan University

Feifan Xiao

Children's Hospital of Fudan University

Bingbing Wu

Children's Hospital of Fudan University

Huijun Wang

Children's Hospital of Fudan University

Wenhao Zhou

Children's Hospital of Fudan University

Xinran Dong (✉ xrdong@fudan.edu.cn)

Children's Hospital of Fudan University <https://orcid.org/0000-0001-9868-8795>

Research

Keywords: Cystic fibrosis, screening panel, prevalence estimation, allele frequency comparison, haplotype construction and comparison

Posted Date: November 19th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-1074469/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Abstract

Background

Cystic fibrosis (CF) is a common, life-threatening genetic disease in Caucasians but rarely reported in Chinese population. The prevalence and population-specific genetic spectrum of CF in China needs to be systematically estimated and compared with Caucasians.

Materials & Methods

We reviewed 30,951 exome-sequencing samples, including 20,909 children samples and 10,042 parent samples, from Chinese Children's Rare Disease Genetic Testing Clinical Collaboration System (CCGT). After the in-lab filtration process, 477 candidate variants of *CFTR* gene were left and 56 variants were manually curated as pathogenic/likely-pathogenic (P/LP). These P/LP variants were adopted to estimate CF prevalence in three methods: the carrier frequency method, the permutation-combinations method and the Bayesian framework method. Allele frequencies of the 477 *CFTR* variants were compared with non-Finland European (NFE) and East Asian (EAS) from gnomAD database. To investigate the haplotype structure difference of *CFTR*, another 2,067 whole-genome-sequencing samples from CCGT and 195 NFE from 1000 genome project were analyzed by Shapeit4 software.

Result

With the 56 manually curated P/LP variants in *CFTR* gene, the estimated Chinese CF prevalence is approximately 1/106,400. Only 21 (37.5%) of the 56 variants were included in Caucasian specific CF screening panels, resulting in significantly under-estimation of CF prevalence in our children cohort (1/114,659 vs. 1/1,121,017, $P=9e-27$) and parents' cohort (1/96,968 vs. 1/797,207, $P=1e-10$). The allele frequencies of six pathogenic variants (G970D, D979A, M469V, G622D, L88X, c.1766+5G>T) were significantly higher in our cohorts compared with gnomAD-NFE population (all P -value<0.1). Haplotype analysis showed more haplotype diversity in Chinese compared to Caucasians. In addition, G970D and F508del were founder mutation of Chinese and Caucasians with two SNPs (rs213950-rs1042077) identified as related genotype in exon region.

Conclusions

Chinese population showed significantly different genetic spectrum pattern in *CFTR* gene compared with Caucasian population, and thus a Chinese-specific CF screening panel is needed.

Introduction

Cystic fibrosis (CF) is an inherited autosomal recessive disease that threatens the patients' whole life. Previous studies found that CF is more common in Caucasian population than in other populations [1]. The preference of CF is approximately 1 in 3,000 for Caucasians, 1 in 4,000-10,000 for Latin Americans and 1 in 15,000-20,000 for African Americans [2, 3]. In the United States, CF occurs in approximately 1 in 400 newborns [4]. However, the reported CF prevalence is always much lower in Asian countries despite that it varies widely from 1:2,560 to 1:350,000 live births [5–7].

The epidemiology of CF has not been well studied in Chinese population. Most published studies focused on the genetic and clinical characteristics of CF in Chinese patient populations. Chinese CF patients have been shown to have novel and different frequencies of the CF transmembrane conductance regulator (*CFTR*) gene variants, which suggests that CF in Chinese population may have a different spectrum of variants comparing with Caucasian population [8, 9]. For example, Gly970Asp was reported as a hot spot in Chinese population while it was not common in Caucasian population and not included in Caucasian screening panels [8]. Therefore, CF screening panels for Caucasian population might not be suitable for Chinese population. What's more, because newborns in China are not screened for CF, potential patients with CF are not systematically identified and CF may be underreported in China.

In this study, we retrospectively analyzed next-generation-sequencing samples in the Chinese Children's Rare Disease Genetic Testing Clinical Collaboration System (CCGT), which is one of the largest genetic databases of the Chinese pediatric population[10]. Then we applied three different methods to estimate CF prevalence in Chinese population and presented quantitative evidence of how

Caucasian CF screening panel is not suitable for Chinese. Furthermore, we systematically compared allele frequencies and haplotype structures between Chinese and Caucasian populations to demonstrate the genetic spectrum differences. Based on these results, we established the panel of CF genetic screening and diagnosis for Chinese population and explained the differences of *CFTR* gene characteristics between Chinese and Caucasian populations.

Results

Estimated CF prevalence of Chinese population is lower than Caucasian population

Totally, we enrolled 20,909 pediatric patients as children cohort and 10,042 parental samples as parent's cohort (Figure 1). After filtration and manually quality assessment for *CFTR* variants in this two cohort, 56 P/LP variants were identified (**Supplementary Table 1**). In the children cohort, the affected frequency of CF was ranged from 1/114,659 to 1/120,525. In the parent's cohort, the estimated CF frequency was ranged from 1/96,968 to 1/105,335 (Table 1). The average estimated prevalence of Chinese CF was around 1/106,400, much lower than in Caucasians (1 in 3,000) and other populations (Latin Americans: 1 in 4,000-10,000, African Americans: 1 in 15,000-20,000) [2, 3].

Table 1
Basic information for the children cohort and parent's cohort with estimated affected frequency by three methods.

	Children cohort	Parents cohort
Total number	20909	10042
Gender (Male/Female)	12776/8113	5009/5047
CES/WES	15871/5038	334/9708
Homozygous or Compound heterozygous	4 (2 Males, 2 Females)	0
Heterozygous	115 (76 Males, 39 Females)	64 (40 Males, 24 Females)
Carrier frequency	-	1/157
Couple's carrier risk	-	1/24619
Method 1: Estimated affected frequency by carrier frequency	-	1/98478
Method 2: Estimated affected frequency by permutation-and-combination	1/120525	1/105335
Method 3: Estimated affected frequency by Bayesian framework (95% confidence interval)	1/114659 (1/167257~1/82456)	1/96968 (1/165913~1/62132)
Average estimated prevalence (mean value of the above five frequencies)	1/106400	

CF screening panels for Caucasians underestimate CF prevalence in Chinese

We treated the identified 56 P/LP variants as a Chinese-specific CF screening panel. Based on this panel, we retrospectively identified three CF patients (**Supplementary Figure 1**). Patient 1 was a 10-year-old boy with two compound heterozygous pathogenic variants (c.935_937delTCT/p.F312del and c.2052dupA/p.Q685fs). Both variants were annotated as DM in HGMD. F312del was inherited from the patient's mother, and Q685fs was a *de novo* variant. Patient 1 was diagnosed as CF with clinical phenotypes of hepatic cirrhosis and hepatosplenomegaly. Patient 2 was a 12-year-old boy diagnosed with hepatosplenomegaly and increased serum hepatic transaminase. A homozygous splicing variant (intron5: c.579+4_579+5delTGinsCA) in patient 2 was identified by CES, where the main effect on splicing was from c.579+5G>A (c.579+4T>C along was marked as VUS). Sanger sequencing found that the homozygous splicing variant was inherited separately from his parents. Both two patients had a negative family history of CF. Patient 3 was an 11-year-old girl with bronchiectasis and recurrent pneumonia. *Pseudomonas aeruginosa* was found in the sputum culture test. A homozygous stop-gained variant (p.L88X) was detected in patient 3 by WES and confirmed by Sanger sequencing that the homozygous variant was inherited from the patient's parents. Besides, we identified another patient (c.1159_1160delTT, c.1242_1243insAACAAAC) without any typical phenotypes, waiting for follow-up interview.

We reviewed typical CF screening panels applied in Caucasians and summarized 140 *CFTR* variants as a Caucasian-specific CF screening panel (**Supplementary Table 2**). We compared the Caucasian-specific CF screening panel and the Chinese-specific panel, and found only 21 variant was shared, which indicated the distinct genetic background in the two populations. Then we applied these two CF screening panels to estimated CF affected frequency in our cohorts and other populations from gnomAD database with Bayesian framework method (Figure 2). The results showed that Caucasian-specific CF screening panel detected much higher affected frequencies in NFE, FIN, AMR and SAS populations, and lower in EAS and our two Chinese cohorts (all $P < 0.1$). Meanwhile the Chinese-specific CF screening panel detected higher affected frequencies in EAS and our two cohorts than in other populations. Notably, CF prevalence would be significantly underestimated in both Chinese children cohort (OR=9.78, from 1/114,659 to 1/1,121,017, $P = 9e-27$) and parents' cohort (OR=8.22, from 1/96,968 to 1/797,207, $P = 1e-10$) if using the Caucasian-specific screening panel.

Allele frequencies of CFTR variants in Chinese is distinct from Caucasians

To further detect the *CFTR* genetic differences between Chinese and Caucasians, we mapped the 56 P/LP variants to CFTR protein structure and calculated variants allele frequencies (AF) of each protein domain. Seven out of 37 protein-related P/LP variants were located in transmembrane domain 2 (TMD2). The total AF of these 7 variants in our children cohort was 2.04×10^{-3} , which was higher than the AF in other four domains (Figure 3, $OR \geq 1.5$), while most variants in Caucasians located in NBD1 domain (**Supplementary Figure 2**). Besides, top two frequent P/LP variants (G970D and D979A) were both located in TMD2 in our Chinese cohorts. These results indicate that TMD2 may be the most important disease-related domain for Chinese population. We also compared the AF of the P/LP variants for three mutation types in different populations (**Supplementary Figure 3**). Four missense variants (G970D, D979A, M469V, G622D), one nonsense variant (L88X) and one splicing variant (c.1766+5G>T) had significantly higher AF in our two cohorts than in gnomAD-NFE (all $P < 0.1$). Three missense variants (F508del, R117C, R117H) had significantly lower AF in our population (all $P < 0.1$).

From the 421 variants curated as VUS or benign in our cohorts, 116 variants were reported as DM in HGMD or P/LP in ClinVar database. After comparing the AF of these conflicting variants among children cohort, parent's cohort, gnomAD-EAS, and gnomAD-NFE, we found that sixteen variants (I125T, N186K, E217G, N287K, K411E, I556V, F650L, E681V, c.2620-97C>T, S895N, R1070Q, R1097C, T1220I, c.3717+45G>A, Q1352H, R1453W) had significantly higher AF (all $P < 4e-7$ and $OR > 10$) in our children and parents' cohorts (**Supplementary Figure 4**). For example, Q1352H has AF 6.5×10^{-3} in our children cohort and 6.4×10^{-3} in our parent cohort compared with no-reporting in gnomAD-NFE ($P < 1e-100$ and $OR = Inf$).

In addition, we compared the AF of polymorphism sites around and on *CFTR* gene (± 1 Mb region) with 2,067 CCGT WGS cohort and gnomAD-NFE, as most SNPs located out of the capture region of exome-sequencing kits. Seventy-eight intron-SNPs had higher minor allele frequency (MAF) in CCGT WGS cohort, while ten intron-SNPs and two linked exon-SNPs (rs1800136, rs1042180, AF 0.235 vs 0.013) had significantly higher MAF in gnomAD-NFE. The SNPs pattern of CCGT was different from NFE, but similar with EAS (**Supplementary Figure 5**).

Haplotype analysis indicated more haplotype diversity in Chinese population

To explore the underlying mechanism of different *CFTR* genetic spectrum between Chinese and Caucasian populations, we analyzed the haplotype pattern based on WGS data. As gnomAD does not provide individual genotype data, we applied the 2,067 WGS cohort from CCGT for haplotype structure construction compared with 195 NFE and 298 EAS from 1000 genome WGS database. Among the

three WGS cohorts, five shared haplotype blocks were detected (Figure 4A). The haplotype construction of the five shared blocks were significantly different between 1000genome-NFE and 1000genome-EAS (all $P < 5 \times 10^{-16}$), and different between 1000genome-NFE and CCGT-WGS cohort (all $P < 5 \times 10^{-20}$), while 1000genome-EAS and CCGT-WGS were only significantly different in the first three blocks (all $P < 1 \times 10^{-4}$) (Figure 4B). When combining the five blocks together, the most frequent haplotype consists 52.05% of 1000genome-NFE, much higher than 36.47% for 1000genome-EAS and 29.22% for CCGT-WGS (all $P < 3 \times 10^{-3}$), indicating less *CFTR* haplotype diversity in Caucasians compared with Chinese (Figure 4C).

Different founder mutations and founder genotypes of *CFTR* were detected in Chinese and Caucasian population

There were two exon SNPs located in haplotype block 3, rs213950 and rs1042077 (Figure 5A), allowing us to study the exon-only joint genotype with P/LP variants in our large-scale children and parent exome-sequencing cohorts. Previous study had reported *CFTR* variant F508del was strongly related with joint genotype “A-T” (combination of rs213950-rs1042077) [11]. The frequency of “A-T” genotype in F508del CF patients was much higher than in 1000genome-NFE (OR=10.5, $P = 3.7 \times 10^{-121}$, Figure 5B). This finding was consistent in our CCGT exome sequencing cohorts which had 3 F508del carriers (OR=51.2, $P = 0.0016$). Furthermore, genotype “A-G” was strongly associated with G970D, which was the most frequent pathogenic variant in our children and parent exome-sequencing cohorts (OR=2.28, $P = 8 \times 10^{-4}$, Figure 5B). When taking all the 56 P/LP variants together, genotype “A-G” was significantly over-represented in alleles carrying P/LP variants ($P = 3 \times 10^{-7}$, Figure 5C).

The AF of F508del associated genotype “A-T” is 0.095 in 1000genome-NFE cohort but no more than 0.015 in neither CCGT exome sequencing cohort, CCGT WGS cohort nor 1000genome-EAS cohort (**Supplementary Table 4**). The high frequency of genotype “A-T” is consistent with the high frequency of F508del variant in Caucasian population. On the contrary, the AF of G970D associated genotype “A-G” is only 0.269 in 1000genome-NFE cohort but 0.414 in CCGT exome sequencing cohort and 0.414 in CCGT WGS cohort, and 0.364 in 1000genome-EAS cohort. In general, F508del and G970D could be founder mutations in Caucasian and Chinese, while genotype “A-T” and “A-G” of rs213950-rs1042077 could be potential risk genotype for *CFTR* P/LP variants in the two populations respectively.

Discussion

In this study, we provided the estimated prevalence of cystic fibrosis in Chinese population based on a Chinese-specific CF screening panel consisting of manually curated *CFTR* pathogenic or likely pathogenic variants in a large-scale exome sequencing Chinese cohort. We also compared the allele frequencies of pathogenic variants, rare non-pathogenic variants and SNPs between Chinese and Caucasian population to investigate the genetic background difference. We attempted to explain the different *CFTR* genetic spectrum in Chinese and Caucasian by analyzing haplotype structures and detecting founder variants.

The prevalence of CF in Caucasian is reported between 1:3,000 to 1:20,000 [2–4], while the CF incidence of Asia population is already known to be much lower than the Caucasian population as 1:2,560 to 1:350,000 [5–8]. In this study, we estimated the prevalence in a robust way. Firstly, all variants of *CFTR* were curated by three experienced geneticists. Secondly, the prevalence of CF was estimated by three methods. The results calculated by the three methods were similar: the Chinese CF prevalence is ranged from 1:105,335 to 1:120,525. This prevalence is much lower than previous reported data. One reason may be that we use the strictly curated P/LP variants to calculate prevalence. In addition, CCGT consists of pediatric patients’ and their parents. This population is not a naturally gathered healthy people cohort and genotype frequency of CCGT does not correspond with the Hardy-Weinberg equilibration. More accurate and robust CF prevalence could be estimated with a naturally gathered health Chinese population.

Because CF has not been included in the Newborn Screening Plan in China, and there were CF patients clinically diagnosed without *CFTR* mutations [12], the Chinese CF patient population may be underestimated. Nowadays, Chinese CF patients’ genetic diagnosis is based on reported *CFTR* variants, most of which are reported in Caucasian population. The Chinese specific variants are unknown. In this study, we applied the 56 P/LP variants as CF screening panel for Chinese population, which could also be used in clinical diagnosis process. We statistically found that the Chinese CF prevalence would be 10% lower if estimated by Caucasian specific CF screening panel. So, it is essential and inevitable to introduce a Chinese specific CF panel in clinical practice. Though we could not directly provide a definite prevalence value by systematically newborn screening, the population-based statistical prevalence may give a preliminary evidence for the underestimation of CF in Chinese populations.

We described the different characteristics of the *CFTR* gene between Chinese population and Caucasian population. Firstly, the pathogenic variants were enriched in TMDs rather than NBDs in Chinese population, where functions were relative less reported except for drug binding variants [13]. This could partially explain the difference of clinical manifestation of CF between Chinese and

Caucasian patients. The three genetic-diagnosed patients in our study had different phenotypes from previously reported patients with the same disease-causing variants [14–16], making genotype-phenotype matching more complicated. Thus, much more patients from different populations are required to draw solid conclusions about genotype-phenotype matching pattern. Secondly, ninety-nine variants, which had been reported as DM in HGMD or P/LP in ClinVar, were curated as VUS or benign level in our study as they had significantly higher allele frequencies in our two cohorts. This finding is consistent with previous study. For example, the carrier frequency of I556V in Asia population is as high as 4.7% [17]. This uncovers the different allele frequency and incomplete penetrance among different populations. Thirdly, polymorphism sites showed the haplotype structure and content were substantial different between Chinese and Caucasians. The frequency of the most frequent haplotype in Caucasian population (60%~70%) was much higher than in EAS and in CCGT population (40%~50%). The top 5 haplotype combinations accounted for 72% of all haplotypes in Caucasian population, while accounts for 54% in EAS and CCGT population. These demonstrate the lack of haplotype diversity in Caucasian population than in EAS and CCGT population.

Founder mutation could help to explain the lower diversity of haplotype and the high frequency of a certain rare genetic disease in a certain population [18]. Although Chinese and Caucasian populations are large and not isolated, the differences of genetic characteristics still suggest the existence of founder effect. Several studies have reported different founder variants of *CFTR* in various races. F508del was reported to account for 30–88% *CFTR* pathogenic variants in non-Chinese populations [17]. Besides, Pompei et al found that most variants were associated with the M470V allele in several European populations which can help to trace the origin of the V allele [19]. Leung et al reported a founder variant I1023R in southern Chinese populations [20]. In this study, we curated the I1023R variant as VUS according to ACMG guideline. However, we found another variant G970D with the highest allele frequency (36 samples, accounts for 19.2% carriers) could be a founder variant in Chinese population, which consisted with a previous study [21]. Our results would be more solid with more next generation sequencing data of Caucasian CF patients and Chinese CF patients. More accurate risk haplotypes could be found if large-scale individual whole genome sequencing dataset, especially from patient samples, could be available in future.

Conclusions

Our study indicated that the genetic spectrum pattern of *CFTR* gene in Chinese population is significantly distinct from Caucasian population, and thus a Chinese-specific CF screening panel is needed.

Materials And Methods

Collection of Chinese population data

This study was approved by the ethics committees of Children's Hospital of Fudan University (2014-107 and 2015-130). Patients of CCGT database who underwent genetic tests from December 2015 to December 2019 were all included. Counselling was performed by physicians prior to genetic testing. Informed consents were obtained from the parents of patients. In total, a cohort consisted of 16,205 clinical exome sequencing (CES) data and 14,746 whole exome sequencing (WES) data was used for prevalence estimation. CES was performed using the Agilent ClearSeq Inherited Disease Kit. WES was conducted by the Agilent Sureselect All Exons Human V5 Kit. Both tests run on the Illumina HiSeq X10 with 150 bp pair-end sequencing. Another cohort consisted of 2,067 unrelated individuals from CCGT who underwent whole genome sequencing (WGS) was used for SNP allele frequency comparison and haplotype estimation (full database was not published, partial samples could be found in [22] and [23]). WGS was operated using a Clinical Laboratory Improvement Amendments and sequenced on an Illumina NovaSeq 6000 platform with 150 bp pair-end read length. All kits covered the *CFTR* gene region. The designed capture region on *CFTR* of CES and WES were showed in **Supplementary Table 3**. Quality control steps were showed in **Supplementary Figure 6**. Details of the sequencing and analysis can be found in our previously published papers [22, 24, 25].

Collection of Caucasian and other populations data

Variant lists of CF screening panel in Caucasian population were collected from public clinical tests and articles (**Supplementary Table 2**). The allele frequencies (AF) of *CFTR* gene in other populations were downloaded from the gnomAD database (V3.1.2) [26]. Non-Finland European, Finnish in Finland, Admixed American, South Asian and East Asian population in gnomAD were used. Gene annotation was from GENCODE [27] (ENSG00000001626, ENST00000003084) and protein domain information was obtained from pfam [28] (uniprot ID: P13569).

Individual genotype datasets from 1000 genome were downloaded from web site [29, 30]. 1000genome-NFE (non-Finland European population) was the combination of CEU (Utah Residents with Northern and Western European Ancestry), TSI (Toscani in Italia), GBR (British in England and Scotland) and IBS (Iberian Population in Spain). 1000genome-EAS (East Asian population) was the combination of CHB (Han Chinese in Beijing, China), JPT (Japanese in Tokyo, Japan), CHS (Southern Han Chinese), CDX (Chinese Dai in Xishuangbanna, China) and KHV (Kinh in Ho Chi Minh City, Vietnam).

Curation of CFTR Pathogenic variants

After quality control of sequencing data and in-lab automated filtration process[24], 477 *CFTR* variants were detected. These *CFTR* variants were curated by two clinical geneticists back-to-back according to the ACMG guideline[31]. After manually curation, in total 56 *CFTR* variants were identified as P/LP variants (**Supplementary table 1**). The identification of CF patients were made by pulmonary physicians and geneticists together according to a published article[32].

Estimation of CF prevalence

The 30,951 samples were divided into two unrelated sub cohorts, the children cohort (15,871 CES samples and 5,038 WES samples) and the parents cohort (334 CES samples and 9,708 WES samples). We obtained the genotype of each 56 P/LP *CFTR* variants in each cohort and estimated CF affected frequency by three methods. The first one was directly based on carrier frequency calculated in the 10,042 parent individuals. The risk for a CF child was defined as the couple's carrier risk (product of carrier frequency) divided by 4 (for autosomal recessive inheritance model). The second one was based on permutation-and-combination. In this strategy, individual gender was involved in the possibility calculation. The third one was based on Bayesian framework, referred from Schrodin et al [33], where 95% confidence interval could be estimated. The main step for this strategy was to calculate the allele number with at least one of the P/LP variants and the total allele number in the cohort. The third strategy was also adopted to estimate CF prevalence in other populations with gnomAD allele count dataset. When using gnomAD allele counts, we hypothesized that no sample could have more than one pathogenic variant in *CFTR* gene, which was acceptable for a cohort with disease-free samples. Detailed calculation process of the three methods were described in **Supplementary Notes**.

Estimation of CFTR gene haplotypes

Three cohorts, 2,067 WGS samples from the CCGT database, 195 NFE samples and 298 EAS samples from 1000 genome database (v5.20130502) were collected. For each cohort, variants information files of *CFTR* (hg19, +-1Mb) were extracted and merged. For the CCGT WGS cohort, phasing was processed by shapeit4[34]. Then phased variant information files of these three cohorts were transformed into plink format. Only single nucleotide polymorphism (SNP) variants with high allele frequency ($\text{maf} \geq 0.01$) and passed the Hardy-Weinberg equilibrium exact test ($\text{hwe} \geq 0.001$) were used. The haplotype block for each population was calculated by using option `-blocks 'no-pheno-req'`. Finally, we performed haplotype-association test between alleles carrying pathogenic variants and those without pathogenic variants by Fisher's Exact test.

Statistical analysis

All statistical analysis was performed by R version 3.6.1. Student's t-test was used for pairwise numeric vector comparison, realized by `t.test` in R. Chi-square test (χ^2 .test) was used for allele frequency comparison. Multiple-test was adjusted by "bonferroni" strategy.

Declarations

Availability of data and materials statements

The data that support the findings of this study are either included in the article (or in its supplementary files) or available from the corresponding author on reasonable request. The data are not publicly available due to privacy or ethical restrictions.

Acknowledgements

We are very grateful to bioinformatic team members in our lab for data analysis and clinical genetic team members for phenotype analysis. We wish to acknowledge Xiao Wang, Bo Liu, Xuemei Zhao, Yanyan Qian, Yao Wang, Xiaomin Peng and Suzhen Xu.

Ethics approval and consent to participate

This study was approved by the ethics committees of Children's Hospital, Fudan University. Informed consent was signed by the patient's parents at clinics or wards. This study was performed within the guidelines of the Declaration of Helsinki.

Consent for publication

All the patients included signed the consent for publication.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by grants from the National Natural Science Foundation of China (31701138), National Key Research and Development Program (2020YFC2006402), Clinical Research Plan of Shanghai Hospital Development Center (SHDC2020CR6028-002) and Shanghai Municipal Science and Technology Major Project (Grant No. 20Z11900600).

Author information

Affiliations

Children's Hospital & Institutes of Biomedical Sciences, Fudan University, National Children's Medical Center, Shanghai, China

Qi Ni, Wenhao Zhou & Xinran Dong

Center for Molecular Medicine, Children's Hospital of Fudan University, National Children's Medical Center, Shanghai, China

Qi Ni, Xiang Chen, Ping Zhang, Lin Yang, Yulan Lu, Feifan Xiao, Bingbing Wu, Huijun Wang, Wenhao Zhou & Xinran Dong

Division of Neonatology, Children's Hospital of Fudan University, National Children's Medical Center, Shanghai, China

Xiang Chen & Wenhao Zhou

Contributions

Xinran Dong and Wenhao Zhou designed the study. Xinran Dong, Qi Ni and Yulan Lu interpreted medical exome/genome results and conducted analysis on the aggregated large cohort data. Xiang Chen, Ping Zhang, Lin Yang, Feifan Xiao, Bingbing Wu and Huijun Wang collected, supervised, and reviewed the clinical data. Qi Ni and Xiang Chen wrote the original manuscript draft. Xinran Dong and Wenhao Zhou supervised the study, critically reviewed, and revised the manuscript for important intellectual content. All authors reviewed the draft and approved the decision to submit for publication.

Corresponding authors

Correspondence to Wenhao Zhou or Xinran Dong.

References

1. Burgel, P.-R., et al., *Future trends in cystic fibrosis demography in 34 European countries*. European Respiratory Journal, 2015. **46**(1): p. 133–141.
2. O'Sullivan, B.P. and S.D. Freedman, *Cystic fibrosis*. The Lancet, 2009. **373**(9678): p. 1891–1904.
3. Sanders, D.B. and A.K. Fink, *Background and epidemiology*. Pediatric Clinics, 2016. **63**(4): p. 567–584.
4. Farrell, P.M., et al., *Diagnosis of cystic fibrosis: consensus guidelines from the cystic fibrosis foundation*. The Journal of pediatrics, 2017. **181**: p. S4-S15. e1.
5. Singh, M., et al., *Epidemiology and genetics of cystic fibrosis in Asia: In preparation for the next-generation treatments*. Respiratory, 2015. **20**(8): p. 1172–1181.
6. Schrijver, I., et al., *The spectrum of CFTR variants in nonwhite cystic fibrosis patients: implications for molecular diagnostic testing*. The Journal of Molecular Diagnostics, 2016. **18**(1): p. 39–50.

7. Palomaki, G.E., S.C. Fitzsimmons, and J.E. Haddow, *Clinical sensitivity of prenatal screening for cystic fibrosis via CFTR carrier testing in a United States panethnic population*. Genetics in Medicine, 2004. **6**(5): p. 405–414.
8. Guo, X., et al., *Clinical and genetic characteristics of cystic fibrosis in CHINESE patients: a systemic review of reported cases*. Orphanet journal of rare diseases, 2018. **13**(1): p. 1–10.
9. Shen, Y., et al., *Clinical phenotypes and genotypic spectrum of cystic fibrosis in Chinese children*. The Journal of pediatrics, 2016. **171**: p. 269–276. e1.
10. Dong, X., et al., *An automatic diagnostic system for pediatric genetic disorders by linking genotype and phenotype information*. medRxiv, 2021: p. 2021.08.26.21261185.
11. Vecchio-Pagán, B., et al., *Deep resequencing of CFTR in 762 F508del homozygotes reveals clusters of non-coding variants associated with cystic fibrosis disease traits*. Human genome variation, 2016. **3**(1): p. 1–9.
12. Groman, J.D., et al., *Variant cystic fibrosis phenotypes in the absence of CFTR mutations*. New England Journal of Medicine, 2002. **347**(6): p. 401–407.
13. Loo, T.W. and D.M. Clarke, *The transmission interfaces contribute asymmetrically to the assembly and activity of human P-glycoprotein*. Journal of Biological Chemistry, 2015. **290**(27): p. 16954–16963.
14. Macek Jr, M., et al., *Identification of a novel nonsense mutation (L88X) in exon 3 of the cystic fibrosis transmembrane conductance regulator gene in a native Korean cystic fibrosis chromosome*. Human mutation, 1992. **1**(6): p. 501–502.
15. Dörk, T., et al., *Detection of more than 50 different CFTR mutations in a large group of German cystic fibrosis patients*. Human genetics, 1994. **94**(5): p. 533–542.
16. Meitinger, T., et al., *In frame deletion ($\Delta F311$) within a short trinucleotide repeat of the first transmembrane region of the cystic fibrosis gene*. Human molecular genetics, 1993. **2**(12): p. 2173–2174.
17. Bienvenu, T., M. Lopez, and E. Girodon, *Molecular Diagnosis and Genetic Counseling of Cystic Fibrosis and Related Disorders: New Challenges*. Genes, 2020. **11**(6): p. 619.
18. Zlotogora, J., *High frequencies of human genetic diseases: founder effect with genetic drift or selection?* American journal of medical genetics, 1994. **49**(1): p. 10–13.
19. Pompei, F., et al., *Haplotype block structure study of the CFTR gene. Most variants are associated with the M470 allele in several European populations*. European journal of human genetics, 2006. **14**(1): p. 85–93.
20. Leung, G.K., et al., *CFTR founder mutation causes protein trafficking defects in Chinese patients with cystic fibrosis*. Molecular genetics & genomic medicine, 2017. **5**(1): p. 40–49.
21. Tian, X., et al., *p. G970D is the most frequent CFTR mutation in Chinese patients with cystic fibrosis*. Human genome variation, 2016. **3**(1): p. 1–6.
22. Wang, H., et al., *Optimized trio genome sequencing (OTGS) as a first-tier genetic test in critically ill infants: practice in China*. Human genetics, 2020. **139**(4): p. 473–482.
23. Wu, B., et al., *Application of Full-Spectrum Rapid Clinical Genome Sequencing Improves Diagnostic Rate and Clinical Outcomes in Critically Ill Infants in the China Neonatal Genomes Project*. Critical Care Medicine, 2021. **49**(10): p. 1674–1683.
24. Yang, L., et al., *Clinical and genetic spectrum of a large cohort of children with epilepsy in China*. Genetics in Medicine, 2019. **21**(3): p. 564–571.
25. Dong, X., et al., *Clinical exome sequencing as the first-tier test for diagnosing developmental disorders covering both CNV and SNV: a Chinese cohort*. Journal of medical genetics, 2020. **57**(8): p. 558–566.
26. Karczewski, K.J., et al., *The mutational constraint spectrum quantified from variation in 141,456 humans*. Nature, 2020. **581**(7809): p. 434–443.
27. Frankish, A., et al., *GENCODE 2021*. Nucleic Acids Research, 2020. **49**(D1): p. D916–D923.
28. El-Gebali, S., et al., *The Pfam protein families database in 2019*. Nucleic Acids Research, 2018. **47**(D1): p. D427–D432.
29. Clarke, L., et al., *The international Genome sample resource (IGSR): A worldwide collection of genome variation incorporating the 1000 Genomes Project data*. Nucleic Acids Research, 2016. **45**(D1): p. D854–D859.
30. *1000genomes*. Available from:
http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/release/20190312_biallelic_SNV_and_INDEL/

31. Richards, S., et al., *Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology*. *Genetics in medicine*, 2015. **17**(5): p. 405–423.

32. Farrell, P.M., et al., *Diagnosis of cystic fibrosis in screened populations*. *The Journal of pediatrics*, 2017. **181**: p. S33-S44. e2.

33. Schrodi, S.J., et al., *Prevalence estimation for monogenic autosomal recessive diseases using population-based genetic data*. *Human genetics*, 2015. **134**(6): p. 659–669.

34. Delaneau, O., J. Marchini, and J.-F. Zagury, *A linear complexity phasing method for thousands of genomes*. *Nature methods*, 2012. **9**(2): p. 179–181.

Figures

Figure 1

Overall Study design. This study consisted of four parts, and within each part, the systematically comparisons among our two cohorts (representing Chinese population) with public reported cohorts from gnomAD-NFE, 1000genome-NFE (representing Caucasian population) and gnomAD-EAS, 1000genome-EAS (control for east Asian population) were performed: (1) CFTR variant pathogenicity curation. (2) CF retrospectively identification. (3) CF affected frequency estimation. (4) CFTR gene haplotype estimation.

Figure 2

Estimated affected frequency on two screening panels (Caucasian, Chinese) by Bayesian framework. The Venn-graph shows the screening variants intersection between Chinese-specific panel and Caucasian-specific panel. Each cohort has two bars with each showing the estimated affected frequency if only use the pathogenic variants in the panel, and the p-value above shows the significance for difference. The vertical line shows the 95% confidence interval.

Figure 3

Allele frequencies of P/LP variants in different domains of CFTR protein. P/LP variants are mapped to CFTR protein domains. Different color label different domains. The height of each bar represents the allele frequency of each variant in each cohort. The violin plot summarizes the total allele frequency of pathogenic variants in each domain.

Figure 4

Haplotype comparison between CCGT, EAS and NFE. (A) Haplotype structure for 1000genome-NFE, 1000genome-EAS and CCGT-WGS cohorts. The shared blocks with tagged SNPs (vertical line) are the intersected regions of these three cohorts. Only haplotype blocks with length larger than 10kb were remained. Five haplotype blocks in 1000genome-NFE, three blocks in 1000genome-EAS and four blocks in CCGT-WGS cohorts were found, resulting in five shared blocks. (B) The distribution of haplotype frequency for each shared block in three cohorts. For each shared block, top five high frequency haplotypes were shown in color and the rest were combined as "other" in grey. The Sankey ribbon between each of the adjacent blocks showed the haplotype intersection statistics. For example, in 1000genome-NFE, 100% of the top 1 haplotype in the shared block 1 were accompanied by top 1 haplotype in the shared block 2. (C) Pie chart for the haplotype frequencies in the three cohorts. The colored chain rectangles indicated the combined haplotype construction of the five shared blocks for the adjacent sector.

Figure 5

Joint genotype analysis for CFTR gene. (A) Location of two exon-tag SNPs (rs213950, rs1042077) in the shared haplotype block 3. (B) Odds ratio of allele frequencies for different rs213950- rs1042077 joint genotype (column) in each cohort (row). (C) Odds ratio of allele frequencies for different rs213950- rs1042077 joint genotype for alleles carry 56 pathogenic variants in CCGT exome sequencing cohorts.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryFile.doc](#)
- [Supplementarytable1.xls](#)
- [Supplementarytable2.xls](#)
- [Supplementarytable3.xls](#)