Published in final edited form as:
Genet Med. 2020 July ; 22(7): 1191–1200. doi:10.1038/s41436-020-0786-5.

EHR phenotypes associated with genetically regulated expression of *CFTR* and application to cystic fibrosis

Xue Zhong, PhD[1,2,#], **Zhijun Yin, PhD**[3,4], **Gengjie Jia, PhD**[5], **Dan Zhou, PhD**[1,2], **Qiang Wei, PhD**[2,6], **Annika Faucon, BS**[7], **Patrick Evans, PhD**[1,2], **Eric R. Gamazon, PhD**[1,2,8,9], **Bingshan Li, PhD**[2,6], **Ran Tao, PhD**[2,10], **Andrey Rzhetsky, PhD**[5,11,12], **Lisa Bastarache, MS**[3], **Nancy J Cox, PhD**[1,2,#]

[1]Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN

[2]Vanderbilt Genetics Institute, Nashville, TN

[3]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN

[4]Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, TN

[5]Department of Medicine, Institute of Genomics and Systems Biology, University of Chicago, Chicago, IL

[6]Department of Molecular Physiology and Biophysics, Vanderbilt University, Nashville, TN

[7]Human Genetics Graduate Program, Vanderbilt University, Nashville, TN

[8]Clare Hall, University of Cambridge, United Kingdom

[9]MRC Epidemiology Unit, University of Cambridge, United Kingdom

[10]Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN

[11]Committee on Genomics, Genetics and Systems Biology, University of Chicago, Chicago, IL

[12]Department of Human Genetics, University of Chicago, Chicago, IL

Abstract

PURPOSE—The ever-growing electronic health records (EHRs) and biobanks offer unique opportunities to study Mendelian diseases. We described a novel approach to summarize clinical manifestations from patient EHRs into phenotypic evidence for cystic fibrosis (CF) with potential to alert unrecognized patients of the disease.

METHODS—We estimated genetically predicted expression (GReX) of *CFTR* and tested for association with clinical diagnoses in Vanderbilt biobank (N=9142 European descendants with 71 cases of CF). The top associated EHR phenotypes were assessed in combination as a phenotype risk score (PheRS) for discriminating CF case status in additional 2.8 million patients from Vanderbilt University Medical Center (VUMC) and 125,305 adult patients including 25,314 CF cases from MarketScan, an independent external cohort.

[#]Corresponding author: xue.zhong@vanderbilt.edu (615-343-0396) and nancy.j.cox@vanderbilt.edu (615-322-2091).

RESULTS—GReX of *CFTR* was associated with EHR phenotypes consistent with CF. PheRS constructed using the EHR phenotypes and weights discovered by the genetic associations improved discriminative power for CF over the initially proposed PheRS in both VUMC and MarketScan.

CONCLUSION—Our study demonstrates the power of EHRs for clinical description of CF and the benefits of using a genetics-informed weighing scheme in construction of a phenotype risk score. This research may find broad applications for phenomic studies of Mendelian disease genes.

Keywords

Mendelian; cystic fibrosis; CFTR; cis-regulated expression; phenotype risk score

INTRODUCTION

Cystic fibrosis (CF) is a recessive Mendelian disease caused by a spectrum of pathogenic variants in the cystic fibrosis transmembrane conductance regulator (*CFTR*) gene. As one of the most common Mendelian diseases, CF continues to pose challenges due to the highly variable clinical manifestations displayed among CF patients [1]. Part of the variability reflects the spectrum of pathogenic variants in the *CFTR* gene, which differ in impact on disease onset, severity and treatment [2-4]. However, the phenotypic variation in CF cannot be explained by the *CFTR* coding variants alone. A variety of studies have identified variants in other regions of the genome that impact the CF phenotypic variability [5-7]. It remains to know whether regulatory variants modulating the expression of *CFTR* might add to the phenotypic variability. Presumably, regulatory variation of Mendelian genes would cause milder phenotypes, in support of this, genome-wide association studies (GWAS) of common diseases revealed overrepresentation of Mendelian genes among the identified risk loci [8]. On the other hand, regulatory variants can also act to modify (reduce) the deleteriousness of coding variants, as shown in cancers and autism [9].

In this study, we proposed to interrogate the phenotypic consequences of regulatory variants of *CFTR*. The aggregate effects of multiple regulatory variants in a gene were determined by using genotypes to impute genetically regulated expression (GReX) from reference resources such as Genotype-Tissue Expression (GTEx) database [10, 11]. Clinical outcomes of predicted expression of *CFTR* were examined through a Phenome-Wide Association Study (PheWAS [12], an unbiased test of association of a genotype with a range of clinical diagnoses) in BioVU, an academic medical center-based biobank with genotypes linked to electronic health records (EHRs) [13]. Moreover, we evaluated in an independent dataset containing EHRs from 2.8 million patients how well the identified EHR phenotypes in aggregate predicted clinical diagnosed CF.

MATERIALS AND METHODS

Data sources

Data were obtained from Synthetic Derivative (SD), the clinical data warehouse at Vanderbilt University Medical Center (VUMC), and BioVU, the VUMC biobank. All the data were de-identified and our study was classified as “non-human subjects” research by

VUMC Institutional Review Board in accordance with the provisions of Title 45 Code Federal Regulations part 46. The genotype dataset contains genome-wide genotype data from 9142 BioVU participants of European ancestries. The genotypes were imputed and phased into Human Haplotype Reference Consortium reference panel (version r.1.1) [14, 15] using IMPUTE2 [16]. Samples of European ancestry were extracted for analysis based on ancestry principal component analysis (PCA). Our second dataset contains phenotype-only data from 2.8 million patients of SD (excluding the 9142 BioVU participants).

Imputing *CFTR* expressions from genotypes

Expression imputation models were previously trained on the GTEx reference panel (version 2015) [17]. GReX of *CFTR* in each tissue was calculated as a weighted sum of the composite alleles in the prediction model. Of the 20 tissue-specific prediction models available for *CFTR*, we focused on the models with modest prediction performance (i.e., correlation of at least 0.1 between predicted and measured expression), and applied the models to the individual-level genotypes of BioVU samples to calculate GReX. We further used phasing information of the genotype data to impute GReX at haplotype-level (hGReX) for tissue ‘brain hypothalamus’.

PheWAS

PheWAS of the GReX of *CFTR* was performed in each tissue separately via logistic regression, adjusting for age, gender, three principal components of ancestry and arrays/ batches. The binary phenotypes (‘phecode’) were derived from billing codes of electronic health records as described previously [12, 18] with the use of the PheWAS package [19]. Each phecode has defined case, control and exclusion criteria and we required two codes on different visit days to instantiate a case for each phecode. Only phecodes with at least 20 cases were included in analysis. Effect sizes were reported by the beta estimates from the regression.

LD-proxy of DF508

DF508 (*CFTR* p.Phe508del) is a 3-basepair deletion (rs113993960, 7:117199645-ATCT-A) on the 508th codon of the *CFTR* gene. Since DF508 was not directly genotyped in initial genotyping arrays, we used the linkage disequilibrium (LD)-proxy allele to tag it (rs111309367_T, r^2=0.4, D’=1). While D’=1, DF508 is less common than this proxy allele, we have P(proxy=1 | DF508=1)=1 and P(DF508=0 | proxy=0)=1. The latter condition indicates 100% specificity of the proxy allele (i.e., non-carriers of $\mathrm{DF508}_{\mathrm{proxy}}$ are also non-carriers of DF508). The former condition can be used to simply the calculation of sensitivity (of the proxy allele to tag DF508) into a ratio of two allele frequencies (AF):

$$S\mathrm{ensitiviy} = \mathrm{P(DF508 = 1 \mid proxy = 1)} = \frac{\mathrm{P(DF508 = 1\,\&\,proxy = 1)}}{\mathrm{P(proxy = 1)}}$$
$$= \frac{\mathrm{P(proxy = 1 \mid DF508 = 1) \cdot P(DF508 = 1)}}{\mathrm{P(proxy = 1)}} = \frac{1 \cdot \mathrm{AF}_{DF508}}{\mathrm{AF}_{proxy}}$$

With 1 in 2500 new born with an incidence of CF in European descent [20] and DF508 being present on 69% to 76% of cystic fibrosis chromosomes in north American CF patients

[21, 22], we estimated that the allele frequency of DF508 in population of European ancestry is approximately 1.67%. This is derived as follows: proption_of_CF_patients_with_DF508=$p^2 + 2p(0.5p) = 2p^2$, and the proportion_of_CF_patients_with_DF508 also equals to $\frac{1}{2500}(0.7)$. So $2p^2 = \frac{1}{2500}(0.7)$, thus p^2=sqrt(1/2500*0.7)=1.67%. Given an AF of 2% for the proxy allele in (non-Finnish) European descent (gnomAD, gnomad.broadinstitute.org/; haploreg4 pubs.broadinstitute.org/mammals/haploreg/haploreg.php), the sensitivity was estimated ~80% (=1.67%/2%). This implies that carriers of $\text{DF508}_{\text{proxy}}$ are not necessarily also carriers of DF508 – a portion of the homozygotes (heterozygotes) of $\text{DF508}_{\text{proxy}}$ are actually heterozygous (non-) carriers of DF508. We denote this proxy allele as $\text{DF508}_{\text{proxy}}$.

GReX of *CFTR* between carriers and non-carriers of CF pathogenic alleles

In addition to DF508, we interrogated additional CF pathogenic alleles (according to ClinVar [version 2017]) that were covered by our genotype data, collectively denoted as '*other*' CF alleles. Heterozygous carriers of these '*other*' CF alleles were carefully determined as carriers of one of these '*other*' CF pathogenic alleles who neither carry $\text{DF508}_{\text{proxy}}$ nor ii) a diagnosis of CF. The condition ii) was to exclude potential compound heterozygotes who carry CF pathogenic alleles uncovered by our genotyping arrays. We tested for difference in hGReX between heterozygous carriers and non-carriers of i) DF508 and ii) '*other*' CF pathogenic alleles using non-parametric Wilcoxon signed rank test.

Measured expression of *CFTR* in relation to DF508

We examined the measured expression of *CFTR* stratified by the dosage of DF508 using the expression data (RNA-seq) and matched genome sequencing data from GTEx (V8 release). We focused on tissues with an averaged expression level of *CFTR* above a threshold (transcript per million (TPM) ≥ 0.01 in GTEx v7). Gene expressions in each tissue were processed according to [23], including steps of quantile-normalization, adjustment for covariates (gender, platform, first five principal components (PCs), and PEER factors to remove hidden batch effects and other confounders in the expression data), and regression of the expression residuals against the dosage of DF508.

Phenotype risk score construction and performance evaluation

In a dataset ('validation set') that contains EHRs from 2.8 million patients (excluding the 9142 participants of the discovery set) from the SD of VUMC, we constructed and evaluated three phenotype risk scores (PheRSs). The EHR phenotypes and weights used to construct each PheRS ($\text{PheRS}^{\text{mapping}}$, $\text{PheRS}^{\text{assoc}}$, and $\text{PheRS}^{\text{hybrid}}$) were shown in Supplementary Table S2. The weights for $\text{PheRS}^{\text{mapping}}$ were extracted from the original paper [12] based on disease prevalence estimated in VU individuals of European ancestry. Since only the relative values matter for the weights, we normalized the weights to have the sum equal to 1. Both the weights of $\text{PheRS}^{\text{assoc}}$ and $\text{PheRS}^{\text{hybrid}}$ were beta (effect size) values from GReX-phenotype associations and normalized to sum up to 1.

The performance of the PheRSs for differentiating CF cases (defined as having the CF diagnosis code in EHRs) from controls was assessed via logistic regression to obtain the probability of the disease occurrence. Because of the highly unbalanced data (~0.1% of CF cases), we calculated the average precision rate (i.e. the area under precision-recall curve) to measure model performance. Each time, 150,000 patients were randomly selected from the validation set, and the average precision was evaluated for both methods (PheRS^{assoc} vs. PheRS^{mapping}). We repeated this process 10 times and compared the performance.

Evaluation of PheRSs in MarketScan

The MarketScan dabases, owned by IBM Watson Health, are a suite of administrative claims-based databases that comprise inpatient and outpatient claims, medical procedure claims, prescription claims, clinical utilization records, and healthcare expenditures. These data are collected from employers, managed care organizations, health plan providers, and state Medicaid agencies. The covered patient population includes more affluent, privately-insured segments of US society [24, 25]. The MarketScan databases describe over half of US population in terms of comprehensive and high-quality coding of diagnoses, procedures, and drug prescriptions. There have been more than 900 peer-reviewed publications since the launch of these databases in 1995, and this number increases even more rapidly in recent years [26, 27].

Here, in order to further evaluate the proposed PheRSs in this study, we used one of the MarketScan databases—the MarketScan Commercial Claims and Encounters database [28]. This Commercial database contains medical claims, outpatient prescription drug claims, and person-level enrollment information. We identified 25,314 CF cases whose first CF diagnosis appearing in the database were at age of 30 years or older and randomly selected 99,991 non-CF controls that are age- and gender-matched (to the CF cases), out of a total of 151 million unique individuals enrolled in the database during the years between 2003 and 2013.

RESULTS

The workflow of the study is described in Figure 1.

EHR phenotypes associated with genetically determined expression of *CFTR*

Using the expression imputation models previously trained on the GTEx reference panel [17], we estimated tissue-specific GReX of *CFTR* in 10 tissues with modest prediction performance (R^2 of at least 0.01; Supplementary Table S1). Phenome-wide scan of the GReX of *CFTR* was performed in BioVU participants of European ancestry (n=9142). In brain hypothalamus, the GReX was associated with clinically diagnosed cystic fibrosis (P=2.3×10^{-39}). Other top-ranked associations reflect clinical symptoms in respiratory, endocrine and metabolic, and gastrointestinal systems (Table 1). These phenotypes capture key classic features of CF, such as ‘Pseudomonal pneumonia’ (P=1.6×10^{-26}), ‘MRSA pneumonia’ (i.e. Methicillin susceptible pneumonia due to Staphylococcus aureus, P=1.3×10^{-20}), ‘bronchopneumonia and lung abscess’ (P=8.4×10^{-14}), and ‘bacterial pneumonia’ (P=6.2×10^{-12}) for respiratory manifestations; ‘disease of pancreas’ (P=2.1×10^{-17}) and ‘secondary diabetes’ (P=5.0×10^{-9}) for endocrine and metabolic manifestations; and ‘nutritional marasmus’ (low weight in infant/child) (P=1.1×10^{-8}), ‘intestinal malabsorption (non-celiac)’ (P=1.1×10^{-7}), ‘severe protein-calorie malnutrition’ (P=0.0001), ‘failure to thrive in childhood’ (P=1.2×10^{-6}), and ‘lack of normal physiological development’ (P=6.7×10^{-6}) for gastrointestinal manifestations (Table 1). The top associations also include rarer phenotypes such as ‘bronchiectasis’ (P=4.9×10^{-19}), ‘hemoptysis’ (coughing up blood or blood-stained mucus, P=1.6×10^{-5}), as well as common phenotypes including ‘nasal polyps’ (P=2.6×10^{-5}), ‘abnormal sputum’ (P=2.6×10^{-5}), and ‘chronic sinusitis’ (P=8.5×10^{-5}). These milder symptoms are consistent with previously reported symptoms in CF cases with adult onset [29-32].

We denoted the top 20 associated EHR phenotypes (excluding CF diagnosis) detected in hypothalamus collectively as the ‘CF-phenome’ (Table 1). Notably, the direction of association was concordantly negative for these top associations (i.e., risk of symptoms was inversely related to the GReX level of *CFTR* in hypothalamus). Similar phenome associations (but less comprehensive) were also detected in two other tissues (brain hippocampus, heart left ventricle) (Supplementary Table S2).

GReX of *CFTR* captures underlying CF coding alleles

Given that the GReX associations captured CF and many of its clinical manifestations, we asked whether the GReX reflects a genuine effect of regulatory variants independent of coding variants, or mainly captures the coding variants in *CFTR* due to linkage disequilibrium (LD). We first conditioned our analysis on DF508, the most common CF-pathogenic variants in European descent. Since DF508 was not directly genotyped in our genotyping arrays, we used the LD-proxy allele (rs111309367, r^2=0.4, D’=1) that tags DF508 with 100% specificity and ~80% sensitivity (Methods). We denoted this proxy allele as $\mathrm{DF508}_{\mathrm{proxy}}$.

After conditioning on the dosage of $\mathrm{DF508}_{\mathrm{proxy}}$, the association of GReX of *CFTR* (in hypothalamus) with the ‘CF-phenome’ attenuated sharply (Table 1). Indeed, GReX of *CFTR* was correlated with the dosage of $\mathrm{DF508}_{\mathrm{proxy}}$, showing a dosage-dependent trend with respect to $\mathrm{DF508}_{\mathrm{proxy}}$ (Figure 2A). None of the individual SNPs that comprise the GReX in hypothalamus are, however, in strong LD with $\mathrm{DF508}_{\mathrm{proxy}}$ (r^2<0.2) (Supplementary Fig. S1). We hypothesized that it is the combination of the noncoding alleles on haplotypes that effectively capture $\mathrm{DF508}_{\mathrm{proxy.}}$ To investigate this, we decomposed the GReX into the sum of two haplotype-level predicted gene expression (hGReX) assuming an additive model (Methods). With phased genotype data, we observed that in heterozygotes of $\mathrm{DF508}_{\mathrm{proxy}}$ (n=414, excluding CF patients), the haplotype carrying $\mathrm{DF508}_{\mathrm{proxy}}$ almost exclusively (98.7%) had lower hGReX than the other (‘wild type’) haplotype (Wilcoxon signed-rank test P<2.2×10^{-16}; Figure 2B).

We then checked whether the expression reduction was also seen in haplotypes carrying CF alleles other than DF508. There are 16 additional CF alleles (according to ClinVar [version 2017]) covered either by our direct genotyping or genotype imputation (Supplementary Table S3). With the allele frequency ranging from 0.001% to 0.2% in BioVU samples, we observed that individuals either carry zero or a single CF allele. Of the carriers (n=121), a few were positive for $\mathrm{DF508}_{\mathrm{proxy}}$ (n=14) or CF case status (n=4); after exclusion of these individuals, we obtained 103 heterozygous carriers for one of these 16 CF alleles who were without diagnosis of CF (Figure 3A). In these heterozygotes, the haplotype carrying a CF-allele on average had lower hGReX than the ‘wild-type’ haplotype (Wilcoxon signed-rank test; $P<4.7\times10^{-12}$; Figure 3B), similar to the observation of $DF508_{proxy}$. In contrast, the load of intronic variants was not correlated with the level of hGReX (P=0.8).

Measured expression of *CFTR* in carriers of DF508

Using the genome sequencing from more than 800 individuals of GTEx database (release V8), we examined the measured gene expression in relation to DF508 in three tissues (hypothalamus, hippocampus, and heart left ventricle) where CF-phenome were detected. The number of carriers of DF508 is small in all three tissues (6 to 8 heterozygous carriers). In brain hippocampus we detected expression reduction of *CFTR* in carriers of DF508 (Wilcoxon rank sum test; P=0.006), while no difference were detected in the other two tissues (Supplementary Fig. S2), likely due to the better correlation between GReX and the actual expressions in hippocampus (r^2=0.074) than in hypothalamus (r^2 =0.011) or heart left ventricle hypothalamus (r^2 =0.025).

Scoring individuals based on EHR phenotypes identified by GReX associations

Since our GReX-associated phenotypes are consistent with clinical features of CF, we assessed whether these EHR phenotypes can be combined to construct a phenotype score to express how close an individual’s EHRs phenotypes are to clinically diagnosed CF. Earlier attempts have built a phenotype risk score for CF by mapping clinical description of Mendelian diseases to EHR phenotypes and then aggregating the relevant EHR phenotypes into a weighted sum with the weights determined by the inverse prevalence of the phenotypes in EHRs [33]. We denoted this score as $PheRS^{mapping}$ (for the composite EHR phenotypes and weights, see Supplementary Table S4).

Here, we constructed an alternative PheRS for CF: we combined the GReX-discovered ‘CF-phenome’ (20 phecodes, excluding CF diagnosis, phecode 499) using weights informed by the effect size estimates from the GReX-phenotype associations (Methods; Supplementary Table S5). This phenotype risk score, denoted as $PheRS^{assoc}$, scored CF patients higher than controls (Wilcoxon rank sum test; $P<2.2\times10^{-16}$) in samples independent of those used discovering the ‘CF-phenome’ (N=31,537 EUs, with 131 CF cases), validating $PheRS^{assoc}$ as a phenotype score for CF (Methods).

Next, we compared the performance of $PheRS^{assoc}$ to $PheRS^{mapping}$ using de-identified EHRs from 2.8 million patients of VUMC (~0.1% were diagnosed as CF cases), independent of the discovery dataset. The precision and recall rates were compared side by side for the scores for ten iterations, and each time a random sample of 150,000 individuals (EHRs) was selected from the total pool (Methods). For each of the 10 datasets, the average precision rate (i.e. area under the precision-recall curve) of $PheRS^{assoc}$ is better than $PheRS^{mapping}$, ranging 20%~36% for the former as opposed to 3%~12% for the latter (Figure 4; Supplementary Fig. 3; Supplementary Table S6). Consistently, the precision and recall of predicted ‘high-risk’ patients (defined as the top 100 high-scoring individuals) of $PheRS^{assoc}$ were better than $PheRS^{mapping}$ across all 10 iterations (Supplementary Table S7).

Since the number of phecodes used in constructing each PheRS is similar (21 phecodes in PheRS^{mapping} vs. 20 phecodes in PheRS^{assoc}, and four shared phecodes between the two scores), we hypothesized that the different weighting schemes may have contributed to the performance difference. To test this, for the PheRS^{mapping}, we kept the constitutive phenotypes unchanged but replaced the original weights by the weights derived from the GReX-phenotype associations detected in hypothalamus (**Table 4A**), the performance of the resulting PheRS (we denote it as PheRS^{hybrid}) almost tripled in the precision-recall rate, ranging from 11% to 23% (Figure 4B). This indicates that the genetics-informed weights substantially outperformed the prevalence-based weights for predicting case status of CF. In addition, the constitutive codes of PheRS^{assoc} generally have better discriminative power for CF than the codes of PheRS^{mapping}, as indicated by the logistic regression of each code against CF status (affected vs. unaffected) that generated larger odds ratios of the codes of PheRS^{assoc} (Supplementary Table S8).

Applying PheRS^{assoc} to the case presentation of a 47-year-old woman who received diagnosis of CF in adulthood [29], the woman ranked 99.9th percentile for CF among 2.8M VUMC patients (Supplementary Table S9), suggesting the potential of our PheRS^{assoc} to effectively alert possible CF cases with adult onset. As a comparison, the PheRS^{mapping} scoring ranked the same woman as 98th percentile for CF [34]. Case presentations of the woman fit 9 out of 20 phenotype components of PheRS^{assoc}, including sinusitis, cough and abnormal sputum, which were not part of the components of PheRS^{mapping}.

We further evaluated the PheRSs in MarketScan, an independent database that contains national-level electronic health records from nearly half of the US population [28] (Methods). After mapping the ICD codes to phecodes, we applied the scoring algorithms to adults aged 30 years or older (Methods). We found: i) PheRS^{assoc} can distinguish CF cases from non-CF controls (one-sided Wilcoxon-rank sum test, $P<3.2E-249$); and ii) PheRS^{assoc} consistently performed better than PheRS^{mapping} (Supplementary Table S10).

DISCUSSION

In this work, we demonstrate that the genetically regulated expression of a gene (*CFTR*) causing a Mendelian disease can be used as a genetic instrument to identify EHR phenotypes consistent with the Mendelian disease (CF). The associated EHR phenotypes can be combined effectively into a PheRS to summarize the evidence of phenotype overlap with CF. The novel weighting scheme guided by the phenotypic associations enhanced the accuracy of PheRS for predicting CF case status. Given that each family doctor is estimated to encounter 2-3 cases of CF during clinical practice [35], it is important to recognize CF cases in adults whose clinical manifestations tend to deviate from those with early onset. The potential of our PheRS to identify possible CF with onset in adulthood points to the clinical utility of this study. With continuous expansion of EHRs and biobanks, our phenotype risk score will continue to evolve, and may eventually facilitate earlier identification of adult onset of CF.

It has been established that specific cells in lung, ionocytes, a minority cell type in lung, express CFTR proteins leading to the canonical lung phenotypes associated with CF [36, 37]. It is therefore not surprising that we did not detect CF-phenome associations from lung which contains bulk expressions of various cell types. In human brain, hypothalamus is the first site of brain discovered for *CFTR* expression [38], and only neurons were found to express CFTR proteins [39, 40]. Lineage relationship traces neurons back to intermediate neuronal progenitors (a form of basal progenitor) [41] and basal progenitors are known to also generate ionocytes [39, 40]. We speculate that brain hypothalamus includes a cell type that shares a developmental lineage with ionocytes in lung and that similar such cell types are present in the other tissues for which we see strong associations to CFTR phenotype. This implies that the cell types expressing CFTR in brain hypothalamus, and the other tissues we observed to show strong associations to CF phenotypes also have higher proportions of cells with a potentially related developmental ontology to the lung ionocytes implicated in CF.

We presented a *de novo* approach that simultaneously identifies the components required for a phenotype risk score: clinical phenotypes and their corresponding weights. The constitutive codes of $\text{PheRS}^{\text{assoc}}$ in general have a better discriminative power for CF than the codes of $\text{PheRS}^{\text{mapping}}$. The weights, which are proportional to the effect sizes, reflect the relative importance of each component EHR phenotype on CF diagnosis (b_i/b_{CF}) as they were measured by a common genetic instrument (genetically determined expression). The genetics-informed weights perform better than the prevalence-based weights as the latter does not capture such relational importance to EHR-based CF diagnosis.

Another contributing factor to the improved performance of our *de novo* approach is that our approach exploits the rich and detailed EHR phenotypes. For example, pneumonia is among the clinical description of CF, and was mapped to EHR ‘pneumonia’ (phecode 480). Our association analysis revealed additional forms of pneumonia, such as ‘bacterial pneumonia (phecode 480.1), ‘pseudomonal pneumonia’ (phecode 480.12) and ‘Methicillin susceptible pneumonia due to Staphylococcus aureus’ (phecode 480.13). These pneumonia terms were all more strongly associated with GReX of *CFTR* ($P<7\times10^{-11}$) than the general term ‘pneumonia’ (P=0.02). This indicates that our *de novo* approach circumvents some of the difficulties in mapping clinical description terms to EHR phenotypes which are structured hierarchically.

Our results do not support a causal role of predicted expression of *CFTR* on CF phenotypes. The lowest predicted expression was also seen in controls; but in CF patients, there was an overrepresentation of the low levels of GReX. Additionally, when we repeated the analysis by excluding the 71 individuals with CF diagnosis (the remaining 9071 patients), all the association signals regarding the ‘CF-phenome’ disappeared (data not shown), suggesting the predicted expression of *CFTR* is unlikely an independent or significant contributor to CF phenotypes, at least at these sample sizes. The observed coupling of CF variants with expression-reducing alleles is consistent with the hypothesis that natural selection favors haplotypes whose composite regulatory alleles reduce the functional impact of the deleterious variants [9]. In line with this explanation, the haplotypes harboring a severe CF allele such as DF508 demonstrated a lower GReX than the haplotypes harboring a less severe CF-allele (Figure 3B). In this regard, since the level of GReX co-evolves with the deleteriousness of total underlying CF alleles due to natural selection, the effect size estimates based on GReX in fact capture the impact of underlying CF pathogenic alleles in aggregate.

Finally, the success of our *de novo* approach of building a phenotype risk score of CF relies on several aspects of CF. The number of CF cases in the discovery dataset (71 CF cases out of ~10K European descent) has empowered our genetic association studies to reveal phenotypes that broadly cover clinical manifestations of CF. These EHR phenotypes comprise the basis for building PheRS^{assoc}, with some being highly specific to CF (e.g. MRSA pneumonia). The availability of the CF diagnosis code in EHRs also made it easier to assign CF case status. CF is the most common recessive Mendelian disease in populations with European ancestries, and was diagnosed in ~0.1% of the patient population of our validation dataset that contains ~2.8 million patients. While it is unclear that CF results can be extended to rarer recessive Mendelian diseases, we believe such investigations may have value for more automated identification of patients with undiagnosed Mendelian diseases and for more complete cataloging of EHR-based phenotypic descriptions of Mendelian diseases.

The study had several limitations. First, the PheRS construction used phecodes derived from ICD billing codes. Although ICD billing codes are ubiquitous and easily shared across health systems, the mapping task from ICD codes to phecodes is not trivial and rather a growing burden. As the massive EHR data continue to accumulate, PheRS constructed using ICD codes directly would simplify the process to adopt PheRS in another health system. Second, there are correlations among the constitutive codes of PheRS that have not yet been systematically handled in the development of PheRS. Although the correlations are weak, taking into account the correlation in PheRS can further increase its performance. Third, there are individuals without cystic fibrosis who scored high (at population level) due to another disease (e.g. septicemia) when the disease manifestations (e.g. pneumonia, bacterial infection) overlap some of the scoring conditions (Supplementary Tables S11-14). Future development of PheRS may consider a more sophisticated machine learning approach to find a better weighting scheme to alleviate these problems. Finally, our study suggests that PheRS could be a valuable tool to stimulate clinical suspicion of patients who may be affected by CF, however, the ultimate utility of PheRS in clinical practice would require prospective studies for further evaluation.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

This work was funded by the National Institutes of Health (NIH) grants R01MH113362, U01HG009086, R35HG010718, R01HL122712, 1P50MH094267, and U01HL108634-01. A.R. also acknowledges support from the DARPA Big Mechanism program under ARO contract W911NF1410333, the King Abdullah University of Science and Technology (KAUST) and a gift from Liz and Kent Dauten. BioVU and the Synthetic Derivative of Vanderbilt University Medical Center are supported by the National Center for Advancing Translational Science grant UL1TR000445 from NIH; the genotypes in BioVU used for the analyses described were funded by NIH grants RC2GM092618 and U01HG004603.

REFERENCES

1. Farrell PM, White TB, Ren CL, Hempstead SE, Accurso F, Derichs N, Howenstine M, McColley SA, Rock M, Rosenfeld M et al.: Diagnosis of Cystic Fibrosis: Consensus Guidelines from the Cystic Fibrosis Foundation. J Pediatr 2017, 181S:S4–S15 e11. [PubMed: 28129811]
2. Ikpa PT, Bijvelds MJ, de Jonge HR: Cystic fibrosis: toward personalized therapies. Int J Biochem Cell Biol 2014, 52:192–200. [PubMed: 24561283]
3. Rowntree RK, Harris A: The phenotypic consequences of CFTR mutations. Ann Hum Genet 2003, 67(Pt 5):471–485. [PubMed: 12940920]
4. Cutting GR: Cystic fibrosis genetics: from molecular understanding to clinical application. Nat Rev Genet 2015, 16(1):45–56. [PubMed: 25404111]
5. Blackman SM, Commander CW, Watson C, Arcara KM, Strug LJ, Stonebraker JR, Wright FA, Rommens JM, Sun L, Pace RG et al.: Genetic modifiers of cystic fibrosis-related diabetes. Diabetes 2013, 62(10):3627–3635. [PubMed: 23670970]
6. Corvol H, Blackman SM, Boelle PY, Gallins PJ, Pace RG, Stonebraker JR, Accurso FJ, Clement A, Collaco JM, Dang H et al.: Genome-wide association meta-analysis identifies five modifier loci of lung disease severity in cystic fibrosis. Nat Commun 2015, 6:8382. [PubMed: 26417704]
7. Wright FA, Strug LJ, Doshi VK, Commander CW, Blackman SM, Sun L, Berthiaume Y, Cutler D, Cojocaru A, Collaco JM et al.: Genome-wide association and linkage identify modifier loci of lung disease severity in cystic fibrosis at 11p13 and 20q13.2. Nat Genet 2011, 43(6):539–546. [PubMed: 21602797]
8. Chong JX, Buckingham KJ, Jhangiani SN, Boehm C, Sobreira N, Smith JD, Harrell TM, McMillin MJ, Wiszniewski W, Gambin T et al.: The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. Am J Hum Genet 2015, 97(2):199–215. [PubMed: 26166479]
9. Castel SE, Cervera A, Mohammadi P, Aguet F, Reverter F, Wolman A, Guigo R, Iossifov I, Vasileva A, Lappalainen T: Modified penetrance of coding variants by cis-regulatory variation contributes to disease risk. Nat Genet 2018, 50(9):1327–1334. [PubMed: 30127527]
10. Consortium GT: Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 2015, 348(6235):648–660. [PubMed: 25954001]
11. Mele M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, Young TR, Goldmann JM, Pervouchine DD, Sullivan TJ et al.: Human genomics. The human transcriptome across tissues and individuals. Science 2015, 348(6235):660–665. [PubMed: 25954002]
12. Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, Wang D, Masys DR, Roden DM, Crawford DC: PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics 2010, 26(9):1205–1210. [PubMed: 20335276]
13. Roden DM, Pulley JM, Basford MA, Bernard GR, Clayton EW, Balser JR, Masys DR: Development of a large-scale de-identified DNA biobank to enable personalized medicine. Clin Pharmacol Ther 2008, 84(3):362–369. [PubMed: 18500243]
14. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, Kang HM, Fuchsberger C, Danecek P, Sharp K et al.: A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet 2016, 48(10):1279–1283. [PubMed: 27548312]
15. Do R, Willer CJ, Schmidt EM, Sengupta S, Gao C, Peloso GM, Gustafsson S, Kanoni S, Ganna A, Chen J et al.: Common variants associated with plasma triglycerides and risk for coronary artery disease. Nat Genet 2013, 45(11):1345–1352. [PubMed: 24097064]
16. Howie BN, Donnelly P, Marchini J: A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet 2009, 5(6):e1000529. [PubMed: 19543373]
17. Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, Carroll RJ, Eyler AE, Denny JC, Consortium GT, Nicolae DL et al.: A gene-based association method for mapping traits using reference transcriptome data. Nat Genet 2015, 47(9):1091–1098. [PubMed: 26258848]
18. Denny JC, Bastarache L, Ritchie MD, Carroll RJ, Zink R, Mosley JD, Field JR, Pulley JM, Ramirez AH, Bowton E et al.: Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nat Biotechnol 2013, 31(12):1102–1110. [PubMed: 24270849]

19. Carroll RJ, Bastarache L, Denny JC: R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. Bioinformatics 2014, 30(16):2375–2376. [PubMed: 24733291]

20. Dodge JA, Morison S, Lewis PA, Coles EC, Geddes D, Russell G, Littlewood JM, Scott MT: Incidence, population, and survival of cystic fibrosis in the UK, 1968–95. UK Cystic Fibrosis Survey Management Committee. Arch Dis Child 1997, 77(6):493–496. [PubMed: 9496181]

21. Kerem B, Rommens JM, Buchanan JA, Markiewicz D, Cox TK, Chakravarti A, Buchwald M, Tsui LC: Identification of the cystic fibrosis gene: genetic analysis. Science 1989, 245(4922):1073–1080. [PubMed: 2570460]

22. Lemna WK, Feldman GL, Kerem B, Fernbach SD, Zevkovich EP, O'Brien WE, Riordan JR, Collins FS, Tsui LC, Beaudet AL: Mutation analysis for heterozygote detection and the prenatal diagnosis of cystic fibrosis. N Engl J Med 1990, 322(5):291–296. [PubMed: 2296270]

23. Stegle O, Parts L, Piipari M, Winn J, Durbin R: Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nat Protoc 2012, 7(3):500–507. [PubMed: 22343431]

24. Putting Research Data into Your Hands with the MarketScan Databases 2016. *Available at:* http://truvenhealthcom/markets/life-sciences/products/data-tools/marketscan-databases [Accessed 2020 Feb 6] 2016.

25. IBM Watson Health, IBM MarketScan Research Databases 2019. *Available at:* https://wwwibmcom/downloads/cas/4QD5ADRL [Accessed 2020 Feb 6] 2019.

26. Kulaylat AS, Schaefer EW, Messaris E, Hollenbeak CS: Truven Health Analytics MarketScan Databases for Clinical Research in Colon and Rectal Surgery. Clin Colon Rectal Surg 2019, 32(1):54–60. [PubMed: 30647546]

27. Quint J: Health research data for the real world: the MarketScan database. Ann Arbor, MI: Truven Health Analytics Inc 2015.

28. Jia G, Li Y, Zhang H, Chattopadhyay I, Boeck Jensen A, Blair DR, Davis L, Robinson PN, Dahlen T, Brunak S et al.: Estimating heritability and genetic correlations from large health datasets in the absence of genetic data. Nat Commun 2019, 10(1):5508. [PubMed: 31796735]

29. Noroski L, Das S, Hajjar J: Case 40–2018: A Woman with Recurrent Sinusitis, Cough, and Bronchiectasis. N Engl J Med 2019, 380(14):1383.

30. McCloskey M, Redmond AO, Hill A, Elborn JS: Clinical features associated with a delayed diagnosis of cystic fibrosis. Respiration 2000, 67(4):402–407. [PubMed: 10940794]

31. Gan KH, Geus WP, Bakker W, Lamers CB, Heijerman HG: Genetic and clinical features of patients with cystic fibrosis diagnosed after the age of 16 years. Thorax 1995, 50(12):1301–1304. [PubMed: 8553305]

32. Rodman DM, Polis JM, Heltshe SL, Sontag MK, Chacon C, Rodman RV, Brayshaw SJ, Huitt GA, Iseman MD, Saavedra MT et al.: Late diagnosis defines a unique population of long-term survivors of cystic fibrosis. Am J Respir Crit Care Med 2005, 171(6):621–626. [PubMed: 15591474]

33. Bastarache L, Hughey JJ, Hebbring S, Marlo J, Zhao W, Ho WT, Van Driest SL, McGregor TL, Mosley JD, Wells QS et al.: Phenotype risk scores identify patients with unrecognized Mendelian disease patterns. Science 2018, 359(6381):1233–1239. [PubMed: 29590070]

34. Bastarache L, Bastarache JA, Denny JC: Case 40–2018: A Woman with Recurrent Sinusitis, Cough, and Bronchiectasis. N Engl J Med 2019, 380(14):1382–1383.

35. Schram CA: Atypical cystic fibrosis: identification in the primary care setting. Can Fam Physician 2012, 58(12):1341–1345, e1699–1704. [PubMed: 23242890]

36. Montoro DT, Haber AL, Biton M, Vinarsky V, Lin B, Birket SE, Yuan F, Chen S, Leung HM, Villoria J et al.: A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. Nature 2018, 560(7718):319–324. [PubMed: 30069044]

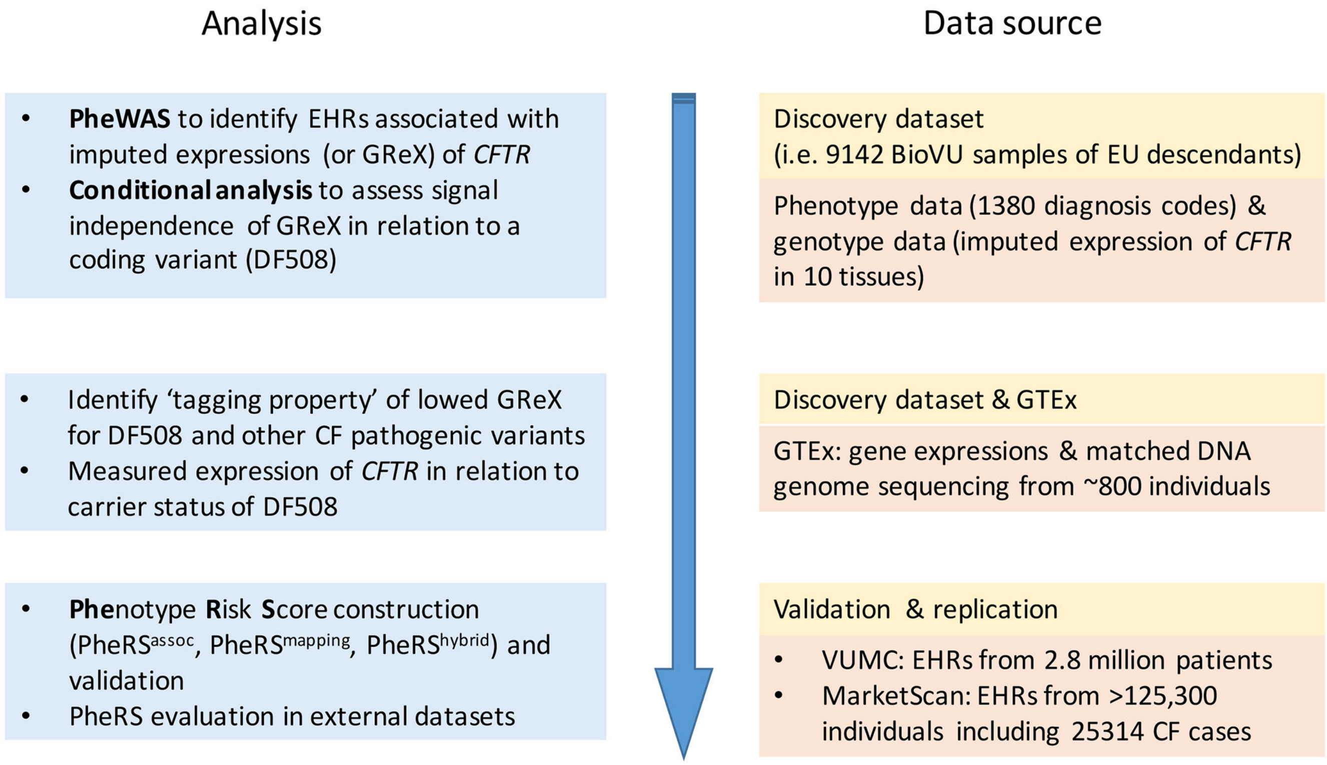
37. Plasschaert LW, Zilionis R, Choo-Wing R, Savova V, Knehr J, Roma G, Klein AM, Jaffe AB: A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. Nature 2018, 560(7718):377–381. [PubMed: 30069046]

38. Mulberg AE, Weyler RT, Altschuler SM, Hyde TM: Cystic fibrosis transmembrane conductance regulator expression in human hypothalamus. Neuroreport 1998, 9(1):141–144. [PubMed: 9592064]

39. Guo Y, Su M, McNutt MA, Gu J: Expression and distribution of cystic fibrosis transmembrane conductance regulator in neurons of the human brain. J Histochem Cytochem 2009, 57(12):1113–1120. [PubMed: 19654104]

40. Marcorelles P, Friocourt G, Uguen A, Lede F, Ferec C, Laquerriere A: Cystic fibrosis transmembrane conductance regulator protein (CFTR) expression in the developing human brain: comparative immunohistochemical study between patients with normal and mutated CFTR. J Histochem Cytochem 2014, 62(11):791–801. [PubMed: 25062999]

41. Kowalczyk T, Pontious A, Englund C, Daza RA, Bedogni F, Hodge R, Attardo A, Bell C, Huttner WB, Hevner RF: Intermediate neuronal progenitors (basal progenitors) produce pyramidal-projection neurons for all layers of cerebral cortex. Cereb Cortex 2009, 19(10):2439–2450. [PubMed: 19168665]



Analysis

- **PheWAS** to identify EHRs associated with imputed expressions (or GReX) of *CFTR*
- **Conditional analysis** to assess signal independence of GReX in relation to a coding variant (DF508)

- Identify 'tagging property' of lowed GReX for DF508 and other CF pathogenic variants
- Measured expression of *CFTR* in relation to carrier status of DF508

- **Phe**notype **R**isk **S**core construction (PheRS^{assoc}, PheRS^{mapping}, PheRS^{hybrid}) and validation
- PheRS evaluation in external datasets

Data source

Discovery dataset
(i.e. 9142 BioVU samples of EU descendants)

Phenotype data (1380 diagnosis codes) & genotype data (imputed expression of *CFTR* in 10 tissues)

Discovery dataset & GTEx

GTEx: gene expressions & matched DNA genome sequencing from ~800 individuals

Validation & replication

- VUMC: EHRs from 2.8 million patients
- MarketScan: EHRs from >125,300 individuals including 25314 CF cases

Figure 1.

Workflow of the study.

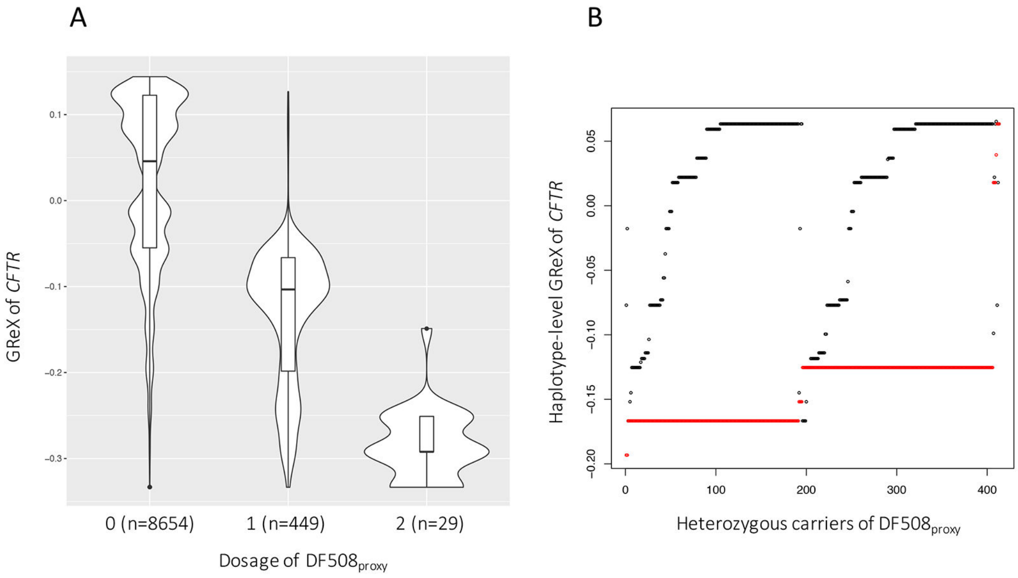
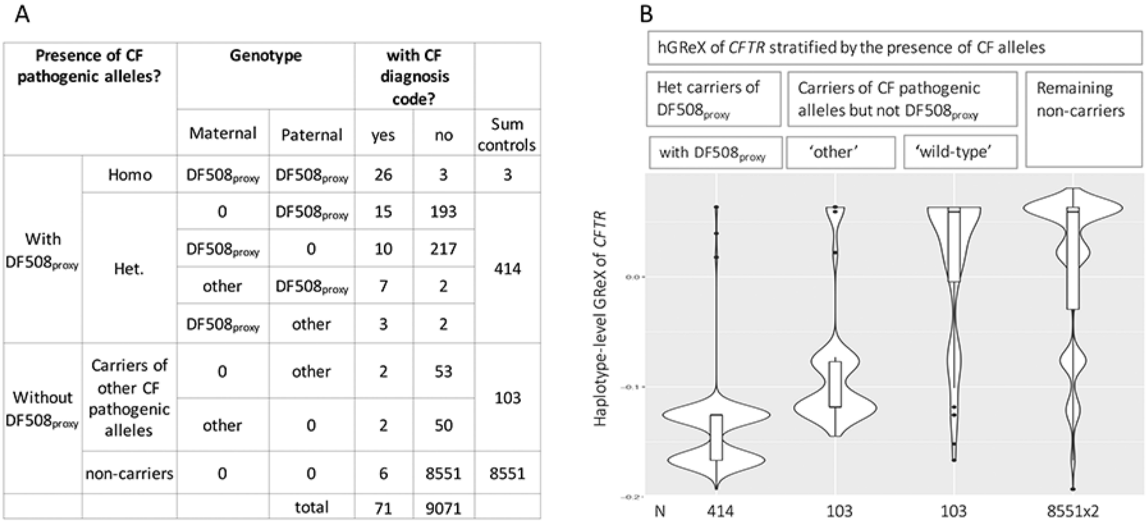


Figure 2.

Genetically regulated expression (GReX) of *CFTR* in brain hypothalamus correlates with dosage of DF508$_{proxy}$. (A) GReX stratified by the dosage of DF508$_{proxy}$. (B) Haplotype-level GReX (hGReX) in heterozygous carriers of DF508$_{proxy}$ who were not diagnosed as CF (n=414). Each heterozygote is represented by a pair of dots, with red referring to the haplotype carrying DF508$_{proxy}$ and black the other ‘wild-type’ haplotype.



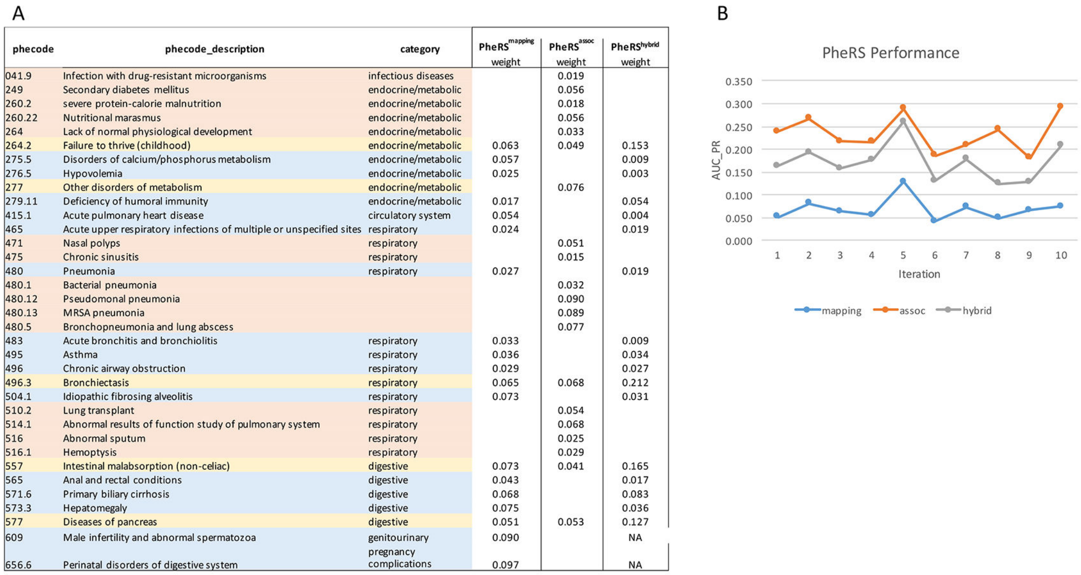
A

Presence of CF pathogenic alleles?		Genotype		with CF diagnosis code?		
		Maternal	Paternal	yes	no	Sum controls
With DF508$_{proxy}$	Homo	DF508$_{proxy}$	DF508$_{proxy}$	26	3	3
	Het.	0	DF508$_{proxy}$	15	193	414
		DF508$_{proxy}$	0	10	217	
		other	DF508$_{proxy}$	7	2	
		DF508$_{proxy}$	other	3	2	
Without DF508$_{proxy}$	Carriers of other CF pathogenic alleles	0	other	2	53	103
		other	0	2	50	
	non-carriers	0	0	6	8551	8551
			total	71	9071	

B
hGReX of *CFTR* stratified by the presence of CF alleles
Het carriers of DF508$_{proxy}$
Carriers of CF pathogenic alleles but not DF508$_{proxy}$
Remaining non-carriers
with DF508$_{proxy}$
'other'
'wild-type'
Haplotype-level GReX of *CFTR*
0.0
-0.1
-0.2
N
414
103
103
8551x2

Figure 3.

Haplotype-level GReX (hGReX) of *CFTR* stratified by the presence of CF alleles. (A) Sample distribution by genotype and CF case status. Case is defined by the presence of CF diagnosis code in EHRs. 'Other' indicates one of 16 CF pathogenic alleles that are also covered by our genotype data. (B) hGReX of haplotypes harboring DF508$_{proxy}$ (n=414), of haplotypes harboring one of 'other' CF-pathogenic alleles (n=103), of 'wild-type' haplotypes from the same carriers (n=103), and of haplotypes from the remaining non-carriers.



phecode	phecode_description	category	$\text{PheRS}^{\text{mapping}}$ weight	$\text{PheRS}^{\text{assoc}}$ weight	$\text{PheRS}^{\text{hybrid}}$ weight
041.9	Infection with drug-resistant microorganisms	infectious diseases		0.019	
249	Secondary diabetes mellitus	endocrine/metabolic		0.056	
260.2	severe protein-calorie malnutrition	endocrine/metabolic		0.018	
260.22	Nutritional marasmus	endocrine/metabolic		0.056	
264	Lack of normal physiological development	endocrine/metabolic		0.033	
264.2	Failure to thrive (childhood)	endocrine/metabolic	0.063	0.049	0.153
275.5	Disorders of calcium/phosphorus metabolism	endocrine/metabolic	0.057		0.009
276.5	Hypovolemia	endocrine/metabolic	0.025		0.003
277	Other disorders of metabolism	endocrine/metabolic		0.076	
279.11	Deficiency of humoral immunity	endocrine/metabolic	0.017		0.054
415.1	Acute pulmonary heart disease	circulatory system	0.054		0.004
465	Acute upper respiratory infections of multiple or unspecified sites	respiratory	0.024		0.019
471	Nasal polyps	respiratory		0.051	
475	Chronic sinusitis	respiratory		0.015	
480	Pneumonia	respiratory	0.027		0.019
480.1	Bacterial pneumonia			0.032	
480.12	Pseudomonal pneumonia			0.090	
480.13	MRSA pneumonia			0.089	
480.5	Bronchopneumonia and lung abscess			0.077	
483	Acute bronchitis and bronchiolitis	respiratory	0.033		0.009
495	Asthma	respiratory	0.036		0.034
496	Chronic airway obstruction	respiratory	0.029		0.027
496.3	Bronchiectasis	respiratory	0.065	0.068	0.212
504.1	Idiopathic fibrosing alveolitis	respiratory	0.073		0.031
510.2	Lung transplant	respiratory		0.054	
514.1	Abnormal results of function study of pulmonary system	respiratory		0.068	
516	Abnormal sputum	respiratory		0.025	
516.1	Hemoptysis	respiratory		0.029	
557	Intestinal malabsorption (non-celiac)	digestive	0.073	0.041	0.165
565	Anal and rectal conditions	digestive	0.043		0.017
571.6	Primary biliary cirrhosis	digestive	0.068		0.083
573.3	Hepatomegaly	digestive	0.075		0.036
577	Diseases of pancreas	digestive	0.051	0.053	0.127
609	Male infertility and abnormal spermatozoa	genitourinary	0.090		NA
656.6	Perinatal disorders of digestive system	pregnancy complications	0.097		NA

Figure 4.

Phenotype risk score (PheRS) construction for CF and performance evaluation. (A) Phecodes and weights used to construct $\text{PheRS}^{\text{assoc}}$, $\text{PheRS}^{\text{mapping}}$ and $\text{PheRS}^{\text{hybrid}}$. Orange and blue indicate phecodes specific to $\text{PheRS}^{\text{assoc}}$ and $\text{PheRS}^{\text{mapping}}$, respectively; gray indicates shared phecodes. $\text{PheRS}^{\text{hybrid}}$ by design has the same constitutive phecodes as $\text{PheRS}^{\text{mapping}}$ with weights derived from GReX associations (NA indicates weights not available due to logistic regression not performed for case number <20). (B) Performance (area under precision-recall curves) across 10 iterations are shown, each with 150,000 patients randomly sampled from a dataset containing de-identified EHRs from 2.8M patients that do not overlap the discovery dataset.

Table 1.

Top associations of EHR phenotypes with GReX of *CFTR* in brain hypothalamus.

	phecode	desscription	category	n_cases	n_controls	Unconditional		Conditioning on DF508	
						beta	p	beta	p
1	499	Cystic fibrosis	respiratory	71	9033	−1.88	2.3E-39	−0.84	3.9E-06
2	480.12	Pseudomonal pneumonia	respiratory	105	6217	−0.95	1.5E-26	−0.20	0.11
3	480.13	MRSA pneumonia	respiratory	82	6217	−0.94	1.3E-20	−0.42	0.001
4	496.3	Bronchiectasis	respiratory	124	6820	−0.71	4.9E-19	−0.19	0.07
5	277	Other disorders of metabolism	endocrine/ metabolic	88	8608	−0.80	1.7E-17	−0.28	0.02
6	577	Diseases of pancreas	digestive	337	8624	−0.43	2.1E-17	−0.17	0.005
7	480.5	Bronchopneumonia and lung abscess	respiratory	71	6217	−0.80	8.2E-14	−0.25	0.08
8	480.1	Bacterial pneumonia	respiratory	385	6217	−0.34	6.2E-12	−0.07	0.26
9	249	Secondary diabetes mellitus	endocrine/ metabolic	80	4936	−0.58	5.0E-09	−0.17	0.20
10	260.22	Nutritional marasmus	endocrine/ metabolic	72	6138	−0.59	1.6E-08	−0.10	0.48
11	510.2	Lung transplant	respiratory	74	7177	−0.57	3.5E-08	−0.26	0.03
12	557	Intestinal malabsorption (non-celiac)	digestive	72	5956	−0.55	1.1E-07	−0.16	0.23
13	264.2	Failure to thrive (childhood)	endocrine/ metabolic	73	6138	−0.51	1.2E-06	−0.07	0.61
14	264	Lack of normal physiological development	endocrine/ metabolic	147	6138	−0.35	6.7E-06	−0.08	0.41
15	516.1	Hemoptysis	respiratory	182	8645	−0.30	1.6E-05	−0.10	0.24
16	471	Nasal polyps	respiratory	49	6193	−0.54	2.6E-05	−0.16	0.31
17	516	Abnormal sputum	respiratory	228	8645	−0.26	2.6E-05	−0.08	0.29
18	514.1	Abnormal results of function study of pulmonary system	respiratory	24	7886	−0.71	6.3E-05	−0.34	0.13
19	475	Chronic sinusitis	respiratory	589	6193	−0.16	8.5E-05	0.02	0.63
20	260.2	severe protein-calorie malnutrition	endocrine/ metabolic	434	6138	−0.18	1.2E-04	−0.09	0.08
21	041.9	Infection with drug-resistant microorganisms	infectious diseases	334	6607	−0.20	1.5E-04	−0.10	0.11

We used a LD-proxy (rs111309367, r2=0.4, D’=1, AF=2%) to tag DF508.

Beta indicates beta per standard deviation (=0.104) of GReX of *CFTR* in brain hypothalamus.