Genomic, transcriptomic, and protein landscape profile of CFTR and cystic fibrosis

Morgan Sanders, James M. J. Lawlor, Xiaopeng Li, John N. Schuen, Susan L. Millard, Xi Zhang, Leah Buck, Bethany Grysko, et al.

Human Genetics

ISSN 0340-6717
Volume 140
Number 3

Hum Genet (2021) 140:423-439
DOI 10.1007/s00439-020-02211-w

Springer

Human Genetics (2021) 140:423–439
https://doi.org/10.1007/s00439-020-02211-w

ORIGINAL INVESTIGATION

Genomic, transcriptomic, and protein landscape profile of CFTR and cystic fibrosis

Morgan Sanders[1] · James M. J. Lawlor[2] · Xiaopeng Li[1] · John N. Schuen[3] · Susan L. Millard[3] · Xi Zhang[4] · Leah Buck[1,5] · Bethany Grysko[6] · Katie L. Uhl[1] · David Hinds[1,2] · Cynthia L. Stenger[5] · Michele Morris[2] · Neil Lamb[2] · Hara Levy[4] · Caleb Bupp[6] · Jeremy W. Prokop[1,7]

Received: 16 May 2020 / Accepted: 25 July 2020 / Published online: 30 July 2020

Abstract

Cystic Fibrosis (CF) is caused most often by removal of amino acid 508 (Phe508del, deltaF508) within CFTR, yet dozens of additional CFTR variants are known to give rise to CF and many variants in the genome are known to contribute to CF pathology. To address CFTR coding variants, we developed a sequence-to-structure-to-dynamic matrix for all amino acids of CFTR using 233 vertebrate species, CFTR structure within a lipid membrane, and 20 ns of molecular dynamic simulation to assess known variants from the CFTR1, CFTR2, ClinVar, TOPmed, gnomAD, and COSMIC databases. Surprisingly, we identify 18 variants of uncertain significance within CFTR from diverse populations that are heritable and a likely cause of CF that have been understudied due to nonexistence in Caucasian populations. In addition, 15 sites within the genome are known to modulate CF pathology, where we have identified one genome region (chr11:34754985-34836401) that contributes to CF through modulation of expression of a noncoding RNA in epithelial cells. These 15 sites are just the beginning of understanding comodifiers of CF, where utilization of eQTLs suggests many additional genomics of CFTR expressing cells that can be influenced by genomic background of CFTR variants. This work highlights that many additional insights of CF genetics are needed, particularly as pharmaceutical interventions increase in the coming years.

Introduction

Cystic Fibrosis (CF) is a rare, autosomal, recessive disorder resulting from mutations within the transmembrane ion transporter *CFTR* (Cutting et al. 1990; Cheng et al. 1990; Zielenski et al. 1991) that impacts ~ 70,000 patients worldwide. CFTR is an anion channel, permeable to both chloride and bicarbonate, regulated by c-AMP (Anderson et al. 1991), with mutations linked to altering multiple molecular outcomes. The homozygous deletion of three DNA bases resulting in the removal of a single phenylalanine (F) at amino acid 508 (known as Phe508del, ΔF508, or delta F508) accounts for around two-thirds of CF cases around the world, with ~ 90% of patients having at least one allele of ΔF508 (Bobadilla et al. 2002). More than 1000 variants have been identified within CFTR in patients with CF, with many of the patients carrying a ΔF508 variant at one allele and either another ΔF508 or a rarer CFTR variant (Bobadilla et al. 2002). In some patients, one or both identified variants are not well defined, and the causal nature is uncertain. These variants are defined as Variants of Uncertain Significance (VUS). While treatment options for CF have grown, there

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00439-020-02211-w) contains supplementary material, which is available to authorized users.

✉ Jeremy W. Prokop
jprokop54@gmail.com

1 Department of Pediatrics and Human Development, College of Human Medicine, Michigan State University, 400 Monroe Ave NW, Grand Rapids, MI 49503, USA

2 HudsonAlpha Institute for Biotechnology, Huntsville, AL 35806, USA

3 Pediatric Pulmonology, Helen DeVos Children's Hospital, Grand Rapids, MI 49503, USA

4 Department of Pediatrics, Division of Pulmonary Medicine, National Jewish Health, Denver, CO 80206, USA

5 Department of Mathematics, University of North Alabama, Florence, AL 35632, USA

6 Spectrum Health Medical Genetics, Grand Rapids, MI 49503, USA

7 Department of Pharmacology and Toxicology, Michigan State University, East Lansing, MI 48824, USA

remains a significant risk of morbidity and mortality, and a resultant need to more fully characterize the wide range of CFTR mutations that give rise to CF. This mission has been particularly taken up by groups such as the HIT-CF, where a computational analysis of CFTR variation can be paired with stem cell technologies of intestinal and lung organoid CFTR function and drug screening.

CFTR is expressed in sweat glands and throughout the respiratory and gastrointestinal tracts. Common variants, like ΔF508, result in the loss of function of CFTR (Crawford et al. 1991; Engelhardt et al. 1992). A large percent of protein variants within CFTR linked to CF alter protein folding or trafficking, resulting in lack of functional protein on the surface of the cell (Cheng et al. 1990) or in gating of the channel (Yu et al. 2012). The targeting of variants in both classes with specialized drugs known as CFTR modulators represent a promising door for many CF patients. While models have been developed for compound heterozygotes of CFTR for sweat chloride levels and pancreatic sufficiency as defined by as defined by fecal pancreatic elastase level (Sebro et al. 2012), additional variant assessments at scale are needed. Having a more robust and rapid VUS prioritization system is one way to begin reclassifying the variants and opening the door for treatment options.

When mutant CFTR is produced by the cell, it is a viable target for pharmacotherapy. Such therapies (referred to as CFTR modulators) have been increasingly available for use by patients since the FDA approval of ivacaftor in 2012 (McKone et al. 2014). Ivacaftor binding restores function of mutations such as G551D, which results in faulty transport of chloride ions by CFTR, altering the mutant CFTR conformation to favor a channel-open state. Ivacaftor monotherapy, however, is indicated only for a small percentage of CF patients due to the rarity of the responding mutations: originally, 3–4% with a G551D allele were eligible, but over time, individual analysis of rare mutations expanded the total indication to ~ 14% of the CF patient population (Feng et al. 2018), demonstrating the therapeutic impact of in-depth knowledge of individual mutations.

Other CFTR mutations, such as ΔF508, require a combination therapy approach, which adds additional compounds to ameliorate additional cellular defects such as protein misfolding and faulty cellular trafficking (Taylor-Cousar et al. 2019). This strategy resulted in the moderately effective combination therapies lumacaftor–ivacaftor (with lumacaftor acting as a folding corrector) (Wainwright et al. 2015), tezacaftor–ivacaftor (tezacaftor, a trafficking corrector) (Taylor-Cousar et al. 2017), which were indicated initially only for ΔF508 homozygotes. The potential of the CFTR-modulation therapeutic approach for ΔF508 was realized in 2019 with the FDA approval of elexacaftor–tezacaftor–ivacaftor "triple combination" therapy, which is indicated for CF patients with at least one ΔF508 allele, approximately 90% of the U.S. patient population (Heijerman et al. 2019; Middleton et al. 2019). As a result of these powerful and genotype-specific treatment options, current clinical guidelines emphasize that all CF patients should undergo CFTR genetic testing to ascertain their genotype (Farrell et al. 2017). However, in patients harboring poorly defined VUS without F508del, the potential utility of the CFTR-modulating therapy remains less well explored, highlighting the need for additional tools that can rapidly and inexpensively screen CFTR variants for potential response to new and existing CFTR modulator drugs and all their potential combinations.

It has been speculated that diverse ethnicities have a prevalence of poorly defined variants within CFTR with early sequence based detection of causal alleles lagging behind in populations such as Hispanics by ~ 30% (Schrijver et al. 2005), giving rise for the need for analysis of diverse CFTR genomic profiling. The need for increased assessments and variant inclusion also in genetic screening stems from the issue of ethnically diverse CFTR variants not being detected in clinical screens. Patients with rare alleles may be reported as false negatives or VUS on both CFTR carrier and diagnostic testing. A rapid analysis tool is in need to find the rare, ethnically unbiased variants as globalization continues to increase and the CFTR variants become more widespread. As the uptake of highly effective modulator therapy among patients with well-defined genotypes increases, ethnic disparities among patients will be magnified by the lack of complete mechanistic understanding of rare genotypes.

While the development of model organisms such as mice (Clarke et al. 1992), pigs (Rogers et al. 2008), and ferrets (Sun et al. 2019) have been able to open the door to define CFTR with promising CF treatment options available, there is evidence that CFTR mutations on different human genomic backgrounds result in different phenotypes (genetic heterogeneity) (Kiesewetter et al. 1993). Clinical heterogeneity in CF is well-described and often independent of CF-causing mutations including ΔF508 homozygous and heterozygous individuals; that is, there remains wide clinical variability among patients with genotypes typically associated with mild or severe disease (Drumm et al. 2012). Indeed, even among ΔF508 homozygotes, individual phenotypes range from severe disease leading to death, transplant during childhood, through moderate disease and survival to geriatric age. These initial insights have given rise to multiple Genome-Wide Association studies (GWAS) to discover the potential genomic modifiers of CF severity (Wright et al. 2011; Blackman et al. 2013; Corvol et al. 2015; Gong et al. 2019). Yet, these insights have been unable to elucidate molecular mechanism of pathology. Thus, the mechanism knowledge of the VUS and genome level modifiers on CF pathology could open many new doors for CF patients. Our goal within this work is to integrate the knowledge

of genomic variants for CF into a systematic informatic sequence-to-structure insight, linking pathogenic variants to ethnically diverse VUS while exploring the mechanisms that modulate CF phenotypes contributed by the rest of the genome.

Methods

Sequence analysis

Sequences for the open reading frame of *CFTR* were extracted from NCBI (www.ncbi.nlm.nih.gov/gene/1080/ortholog/?scope=7776) for vertebrate species. Open reading frames were extracted using TransDecoder (Haas et al. 2013). Sequences were aligned using ClustalW codon (Larkin et al. 2007), removing any sequences with ambiguity or missing exons found in > 90% of the other sequences. Following alignment codons were assessed for selection using dN-dS using a Maximum likelihood Muse-Gaut model (Muse and Gaut 1994) for Tamura-Nei nucleotide substitutions (Tamura and Nei 1993) using HyPhy (Pond et al. 2005) and MEGA (Tamura et al. 2011). With codon selection we performed analysis of each amino acid and for a 21-codon sliding window as done before (Prokop et al. 2017). Post-translational modifications for CFTR were extracted from UniProt (Apweiler et al. 2004). Amino acids for each mouse and pig, common model organisms, were also assessed for each human position. The information was extracted for all amino acids into an amino acid details file (Supplemental Excel file). Phylogenetic analysis was performed using the 233 open reading frame sequences identified for CFTR using maximum likelihood and 1000 bootstraps.

Protein modeling and dynamics

The CFTR protein model from our previous paper (Prokop et al. 2017) with protonation at pH of 7.4 was embedded into a phosphatidyl-ethanolamine (PEA) lipid membrane with water and 0.9% Na/Cl equilibrated on each side of the membrane using YASARA (Krieger et al. 2009). Following energy minimizations of the protein within the lipid membrane, molecular dynamic simulations (mds) was performed for 50 ns using the AMBER03 force field (Duan et al. 2003) followed by analysis with the YASARA md_analyze and md_analyzeres macros (Krieger and Vriend 2015), which included the output of a dynamic-cross correlation matrix (DCCM). The energy minimized 3D structure was saved as a PDB file, loaded into PyMol, color coded, and exported as a VRML2 file. The VRML2 file was uploaded to Shapeways, sized to small or large prints, and made into a product for purchase.

Genomic variant characterization

Genomic variants were compiled for CFTR from the CFTR1, CFTR2 (Castellani and CFTR2 team 2013), ClinVar (Landrum et al. 2016), gnomAD (Lek et al. 2016), TOPmed, and COSMIC (Forbes et al. 2011) databases in October 2019. Variants were extracted from each and integrated together, bringing categories from each database along with our amino acid table data above into a compiled table of variants. All of the variants were assessed with PolyPhen2 (Adzhubei et al. 2010), Provean (Choi and Chan 2015), SIFT (Ng and Henikoff 2003), Align-GVGD (Tavtigian et al. 2006). Included categories are mouse/pig conservation, number of variants in CFTR/CFTR2 databases, CFTR2 allele frequency, CFTR2 annotated % pancreatic insufficient and variant determination, ClinVar annotation, COSMIC count, TOPmed frequency, known rsID, gnomAD population frequencies and max frequency, Polyphen2, Provean, SIFT, Align-GVGD, start/stop annotation, conservation/21-codon scores from above, dynamics correlation amino acids, number of pathogenic annotated variants correlated to each variant, the molecular movement of the amino acid, and the presence of posttranslational modifications. The gnomAD v2.1.1 non-TOPMed samples were used for annotation of population allele frequencies (population sample sizes for the total of 135,727 individuals are African = 10,291, Latino = 17,634, Ashkenazi Jewish = 5068, East Asian = 9956, European Finnish = 12,561, European non-Finnish = 61,378, Other = 3538, and South Asian = 15,308). A simple impact score of each variant was generated by converting each tools prediction into binary (0 = nonfunction, 1 = function) and combined with the conservation score (0–2) and multiplied by the 21-codon conservation. For top VUS of non-Caucasian populations we extracted variants in linkage disequilibrium > 0.8 R^2 using SNiPA (Arnold et al. 2015).

GWAS LD block analyses

The updated Genome-Wide Association Study Database was extracted on November 2019 from the EBI/NHGRI catalog (MacArthur et al. 2017). Variants mentioning "Cystic Fibrosis" were extracted. The rsID for each CF lead SNP was imputed through SNiPA proxy search (Arnold et al. 2015) for all SNPs > 0.8 R^2 in American and European populations using 1000 Genomes, Phase 3 v 5. The SNPs were cleaned to remove all repeats and binned into 15 genomic loci. The SNP list was queried against the category 1–3 variants of RegulomeDB (Boyle et al. 2012) to identify top regulation potential, against PolyPhen2 to identify missense variants, against GTEx eQTL (Lonsdale et al. 2013; GTEx Consortium et al. 2017) lists to identify expression genes, and against the entire EBI/

NHGRI GWAS catalog to identify additional traits. A detailed analysis of chr11:34754985-34836401 was done by visualizing Roadmap Epigenomics annotations (Roadmap Epigenomics Consortium et al. 2015) and ENCODE ChIP-Seq data from K562 and HepG2 (ENCODE Project Consortium 2012). Reads from RNAseq experiments of Caco2 cells (SRA files SRX2169678, SRX2169677, SRX2169676, SRX2169675, SRX2169674, SRX2169673, SRX2169672, SRX2169671, SRX1038553) were mapped using NCBI SRA BLAST against chr11:34754985-34836401 followed by extraction of reads and alignment using Ugene (Okonechnikov et al. 2012).

CFTR expressing cell comodifier analysis

Single cell RNAseq analysis was done by querying CFTR against PanglaoDB (Franzén et al. 2019) and the EBI Single Cell Expression Atlas (Papatheodorou et al. 2020). Detailed analysis of the Tabula muris project single cell RNAseq was performed (Tabula Muris Consortium et al. 2018) for each tissue. The mouse lung expression count table from the Tabula muris project (Tabula Muris Consortium et al. 2018) was normalized to mapped reads per one million reads in each cell line. CFTR expression was binned into cells with > 10 mapped reads and < 10. The Log2 fold change of the two groups as well as the number of cells where each gene was expressed was calculated for all genes followed by GO enrichment and network analysis (Franceschini et al. 2013) or for eQTLs of the top segregating genes using GTEx data (GTEx Consortium et al. 2017).

Results

Building a sequence-to-structure database of CFTR amino acids

A total of 233 vertebrate species sequences of *CFTR* were obtained from NCBI, representing diverse evolution from humans to fish (Fig. 1a). Following alignment and codon selection analysis of the 233 open reading frame sequences (from start codon to stop codon), we generated scores for every amino acid/codon. Of the 1480 human codons, 18 have fixed codons (3 ATG, 1 CAG, 1 GAG, 13 TAG), of which ATG (M) and TAG (W) are single use codons expected to be fixed, while the CAG (codon 996) and GAG (codon 403) suggest unique codon fixation. The conservation score for each amino acid was assessed for linear motif conservation using a 21-codon sliding window additive scoring system, taking each position plus ten before and ten after to score the most conserved linear motifs within CFTR (Fig. 1b). The center of the most selected motifs are labeled on the figure.

Next, we built an integrative structural assessment for CFTR (Fig. 2). The protein model for CFTR was generated for amino acid 1–1480 followed by embedding within a lipid membrane and water added to each side of the membrane (Fig. 2a, b). Throughout a 50 ns (ns) molecular dynamics simulations (mds), the protein reached equilibrium of movement around 1 ns of mds (Fig. 2c), allowing for the calculation of each amino acids' movement throughout the simulation (Fig. 2d). The average Root-Mean-Square Fluctuation (RMSF) for amino acids is 2.7 Å. A total of 91.6% of the amino acids have a RMSF below 5 Å with 42% below 2 Å, which are indicative of well-folded amino acids often contributing to hydrophobic collapse. Only 8.4% of amino acids in CFTR have a

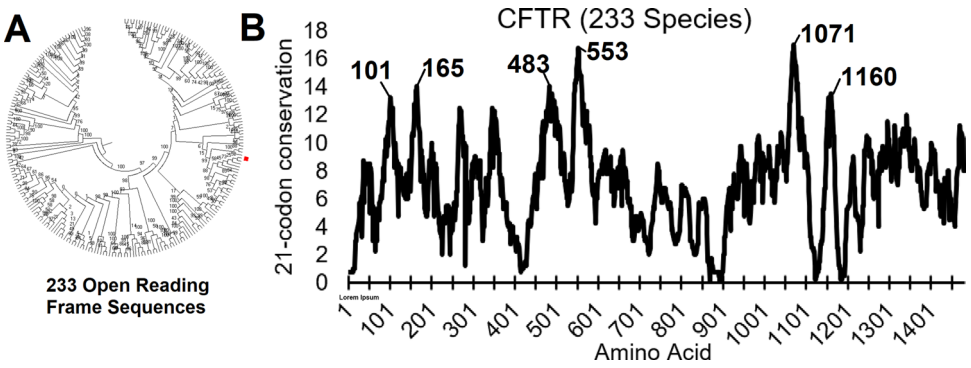


Fig. 1 CFTR Evolution. **a** Phylogenetic tree of 233 species open reading frame (ORF) sequences of CFTR. The red square is the human CFTR sequence. Numbers at each node represent the percent of clustering within 1000 bootstrap analyses. **b** Codon selection and amino acid conservation analysis of the 233 sequences of CFTR placed on a 21-codon sliding window. The center of the top six motifs within CFTR are labeled for human amino acid number

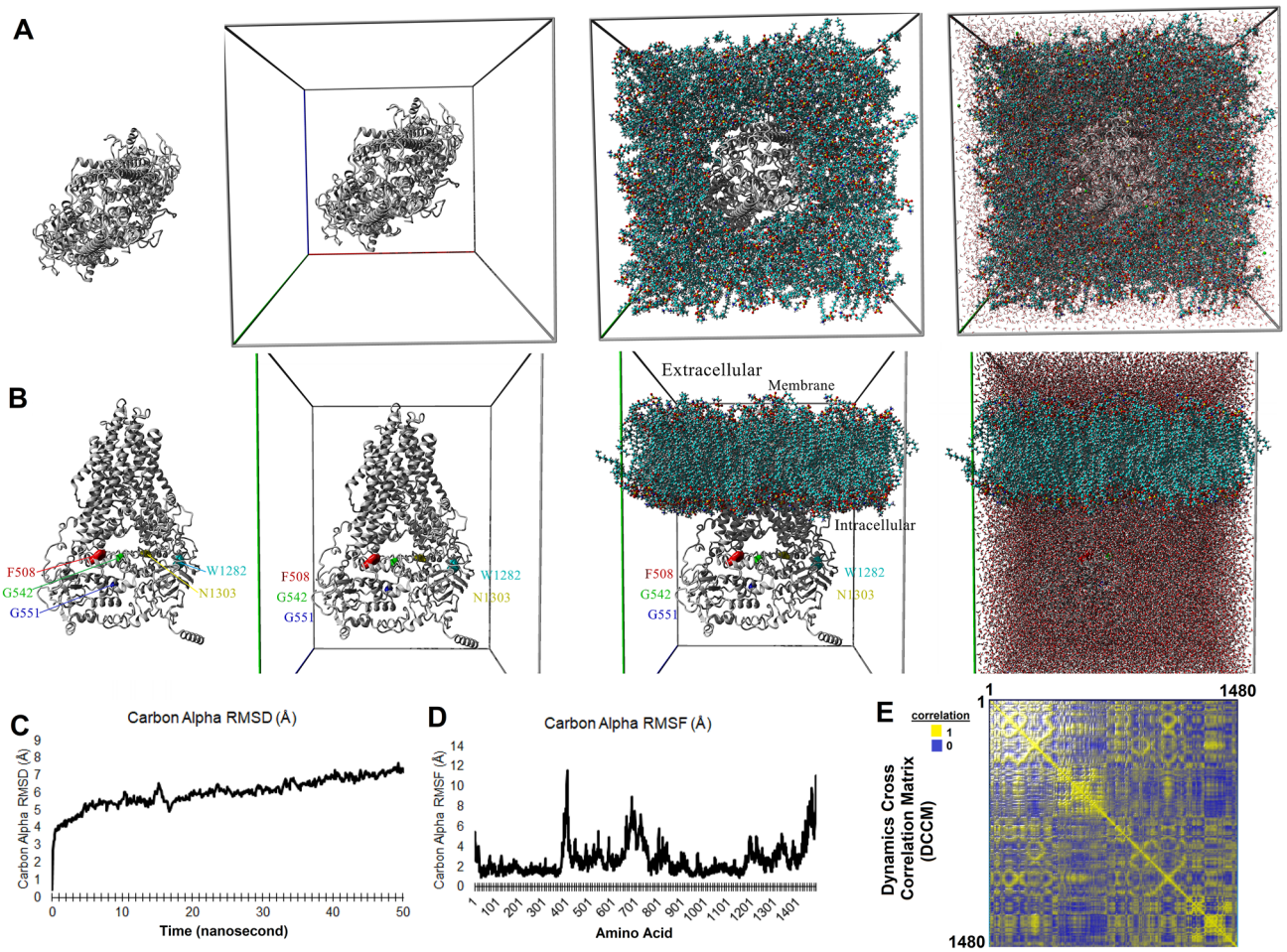


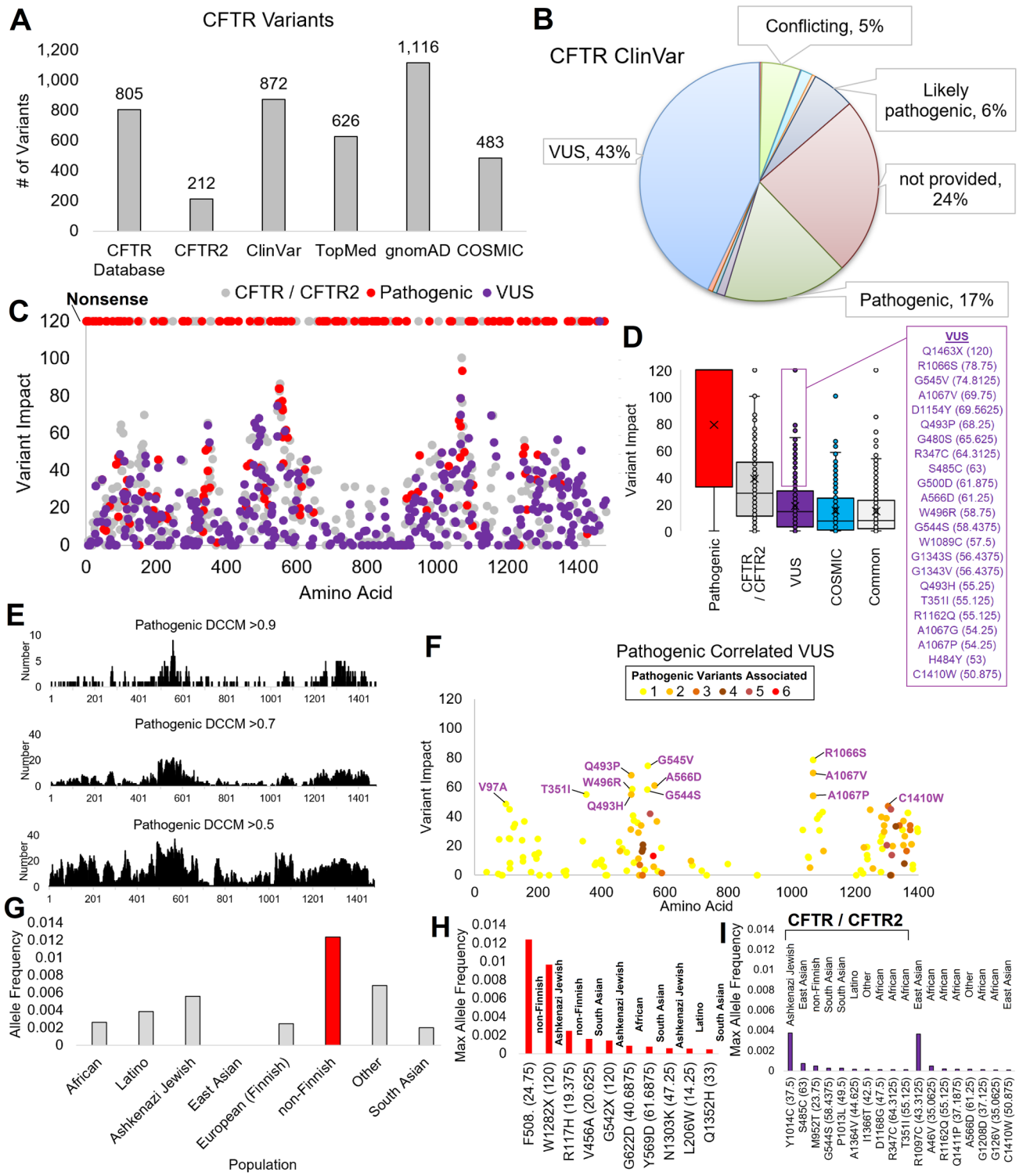
Fig. 2 CFTR structure and dynamics. **a** Top view of CFTR model, model in simulation box, model embedded into lipid membrane, and water added (left to right). **b** Side view correlating to panel A, with amino acids marked for common variants. **c** 50 ns (ns) of molecular dynamic simulations of CFTR protein embedded into a lipid membrane with water on the intracellular and extracellular sides. Data shows the Root-mean squared deviation (RMSD) of the average carbon alpha from the initial structure to each time point of the simulation. **d** The carbon alpha root mean squared fluctuation (RMSF) of each amino acid throughout the 50 ns simulation. **e** Dynamics cross correlation matrix (DCCM) of amino acids. Sites approaching a value of 1 (highly correlated) are in yellow and sites with no correlation in blue

movement above 5 Å, sites that are often loops of the protein domains and the untethered N- and C-terminals.

Utilizing the amino acid trajectories throughout the mds, the correlation of amino acid movement for every amino acid to every other amino acid was calculated for a Dynamics Cross Correlation Matrix (DCCM, Fig. 2e). Each amino acid on average has 84.7 other amino acids correlated > 0.5, 51.2 > 0.6, 29.1 > 0.7, 14.4 > 0.8, and 4.6 > 0.9. This DCCM allows for the connection of amino acids to others, opening a door to build correlations between well-defined variants for CFTR and those variants such as VUS that are less well understood. The compiled sequence, structural and mds analysis for each codon can be found within the Amino Acid Details tab of the supplemental Excel file.

Genomic variant database

Through the integration of the sequence and structural insights at every amino acid, it is possible to build more robust screening platform for genomic insights. The integration of variants from the CFTR1, CFTR2, ClinVar, TOPmed, gnomAD, and COSMIC databases represents the largest consolidated analysis of CFTR variants to date. A total of 2006 unique CFTR variants are present within these databases (Compiled Variants tab of Supplemental Excel File). The gnomAD database contains the largest number of variants (1,116) followed by ClinVar (872) and the CFTR (805) databases (Fig. 3a) at the time of analysis. Dissection of the ClinVar annotations for the 872 variants shows the largest percent to be those of VUS (43%) followed by not provided



(24%) and pathogenic (17%) (Fig. 3b). This means the largest clinical void is the power to call VUS functional or not based on current tools.

All variants within the updated list were assessed using PolyPhen2, Provean, SIFT, Align-GVGD, our conservation score, and our linear motif scores. Many of the ClinVar annotated pathogenic variants of CFTR are nonsense or frameshift variants, while additional variants seen within the CFTR/CFTR2 database and ClinVar annotated VUS tend to be single amino acid changes (Fig. 3c). ClinVar pathogenic variants have the highest predicted variant impact scores of all the variant groups analyzed, with a diverse list

◂**Fig. 3** Integrated knowledgebase of CFTR variants. **a** The number of unique missense, nonsense, or frameshift mutations found within CFTR from various databases. **b** ClinVar annotations for CFTR variants. **c** Variant impact scoring for CFTR variants annotated from ClinVar pathogenic (red), CFTR/CFTR2 databases (gray), or ClinVar VUS (magenta). **d** Box and whisker plots for values in each group of panel C with additional values for COSMIC and gnomAD/TOPmed (common) variants. **e** The number of amino acids correlated to each site of CFTR throughout the molecular dynamic simulations with a cutoff of 0.9 (top), 0.7 (middle) or 0.5 (bottom) correlation. **f** VUS that are correlated in dynamics to pathogenic variants. Color corresponds to the number of pathogenic amino acids associated with each site. **g** The allele frequencies for ΔF508 in various ethnicities of gnomAD with the red box identified as the highest of all populations. **h** The highest allele frequency from gnomAD for pathogenic annotated variants. Each is listed as the variant with the predicted impact score in brackets. **i** The highest allele frequency from gnomAD for VUS annotated as functional from our predication scores. Each is listed as the variant with the predicted impact score in brackets

of VUS sites that overlap with pathogenic scoring levels (Fig. 3d). The variants were annotated such that any listed in the CFTR/CFTR2 database with ClinVar annotation were included in the Pathogenic or VUS groups. A total of 50.5% (104/206) of the pathogenic variants are nonsense with a max score of 120, while 13.7% (61/444) of CFTR/CFTR2 database variants are nonsense. This confirms that pathogenicity analysis of CFTR variants is biased to nonsense variants, even though the most prevalent variant in CFTR, ΔF508, does not fall within this group. Therefore, amino acid assessments of CFTR are critical to define many of the VUS, with our amino acid impact scores allowing us to prioritize VUS of most potential impact (Fig. 3d). The VUS list was further filtered through our mds data by annotating all of the amino acids that correlate in movement to pathogenic variants sites (Fig. 3e), defining 13 high-impact, pathogenic associated VUS including R1066S, G545V, A1067V, Q493P, A566D, W496R, G544S, Q493H, T351I, A1067G, A1067P, C1410W, V97A (Fig. 3f).

Genomic inheritance of functional variants

Heritability of functional variants is of prime importance for developing robust ethnically diverse screening platforms for CF. Thus, we assessed our genomic variant list through allele frequencies to define variants that are potentially heritable. Beginning with ΔF508, the most defined functional variant within CFTR, it is believed the origins of the variant is very old (Vecchio-Pagán et al. 2016). The ΔF508 is found in all populations except for that of East Asian from the gnomAD database, with the highest prevalence in European non-Finnish individuals (Fig. 3g). The widespread occurrence of ΔF508 suggests an ancient origin, being maintained at a constant low frequency throughout human migration and recent mixing of populations. Like ΔF508, additional pathogenic variants are seen across ethnicities including W1282X, R117H, V456A, G542X, G622D, Y569D, N1303K, L206W, and Q1352H (Fig. 3h). From the pathogenic variant annotations of ClinVar, 52.9% (109/206) of variants are observed within gnomAD database, conferring autosomal recessive inheritance of familial variants for a large portion of CF and suggesting a role of de novo variant formation in some individuals.

Yet, 65.4% (242/370) of the ClinVar annotated VUS are also found within gnomAD, making it challenging to use inheritance to rule out functionality as done with many rare diseases and disorders. Using our filtering for functional VUS, we identified ten variants found within the CFTR1 or CFTR2 database that follow diverse population inheritance at low allele frequencies with an additional eight that were not present within either database (Fig. 3i). A total of eight of the VUS (I1366T, T351I, R1097C, G544S, Q1411P, A566D, G126V, C1410W) have mds correlation to pathogenic variants, further suggesting their connection to CF pathology. Five of the VUS (I1366T, T351I, R1097C, Y1014C, S485C) have known linkage disequilibrium inheritance blocks within the 1000 genomes data and are present in populations outside of Finish or non-Finish European. Overall, this would heavily support a functional inheritance role of three CFTR VUS (I1366T, T351I, R1097C) found within diverse ethnicities (Other, African, and East Asian) that have been overlooked for CF involvement and screening. Details for all 18 heritable and likely functional VUS can be found within Table 1.

Defining CF genomic modifiers

While defining the variants for CFTR linked to CF, there is knowledge that additional genetics can influence CF pathology and outcomes. Therefore, we sought to assess these sites, detailing one site overlapping CF and COPD pathology. Taking the lead SNPs for CF within the GWAS catalog (Wright et al. 2011; Blackman et al. 2013; Corvol et al. 2015; Gong et al. 2019) identified 15 linkage disequilibrium (LD) blocks, with 4 of them having multiple lead SNPs (Table 2). All variants with > 0.8 R^2 correlation with the lead SNPs were assessed for functional missense variants, gene regulation potential using RegulomeDB, expression quantitative trait loci (eQTLs), and assessment for additional trait associations from the GWAS catalog returning the top finding in Table 2.

In total, we identify 754 LD SNPs, two of which are missense, MUC4 S585A and AHRR D627H, both predicted to be nonfunctional in multiple tools. RegulomeDB allows an assessment of transcription factor-binding and chromosome annotation states for any variant. One variant, rs9271589 (from LD block chr6:32460285-32644258) had a score of 1b, the top identified potential alteration of gene regulation. This variant is predicted to alter a ZNF628-binding

Table 1 Ethnically diverse and functional predicted variants of uncertain significance (VUS) in CFTR

Variant	CFTR database	CFTR2 Database	COSMIC Database	Max Freq	Max Freq population	Poly-Phen2	Provean	SIFT	Align-GVGD	Conservation	21-Codon conservation	DCCM > 0.9	Pathogenic DCCM > 0.9	RMSF (Å)	LD SNPs
rs200955612 (I1366T)	1	0	0	0.000163026	Other	1	1	1	1	1	8.5	6	1	1.77	12
rs1800086 (T351I)	1	0	0	0.000120163	African	1		1	1	1.5	12.25	3	1	1.172	7
rs201591901 (R1097C)	0	0	1	0.003660983	East Asian	1	1	1	1	1.25	8.25	1	1	1.725	7
rs149279509 (Y1014C)	1	6	0	0.003763026	Ashkenazi Jewish	1	1	1	1	1	7.5	1	0	1.745	3
rs138427145 (S485C)	1	0	0	0.00075188	East Asian	1	1	1	1	1.25	12	0	0	2.632	3
rs142773283 (M952T)	1	0	0	0.000503338	non-Finnish	1	1	1	1	1	4.75	1	0	1.456	0
rs762224063 (G544S)	1	0	0	0.000261455	South Asian	1		1	1	1.25	13.75	7	1	3.746	0
rs193922516 (P1013L)	1	0	0	0.000261352	South Asian	1	1	1	1	1.5	9	0	0	1.229	0
rs397508670 (A1364V)	1	0	2	0.000173541	Latino	1	1	1	1	1.25	8.5	3	0	1.868	0
rs150326506 (D1168G)	1	0	0	0.000123107	African	1	1	1	1	1	9.5	1	0	1.897	0
rs397508147 (R347C)	1	0	1	0.000123031	African	1	1	1	1	1.25	12.25	0	0	1.269	0
rs151020603 (A46V)	0	0	0	0.000480769	African	1		1	1	1.25	8.25	1	0	1.313	0
rs1800120 (R1162Q)	0	0	0	0.000200353	African	0.5	1	1	0.5	1.5	12.25	3	0	1.268	0
rs150177304 (Q1411P)	0	0	0	0.000184615	African	1		1	1	1.25	8.75	5	2	2.91	0
rs1375786834 (A566D)	0	0	0	0.000163881	Other	1	1	1	1	1	12.25	4	2	1.871	0
rs746103666 (G1208D)	0	0	0	0.000123077	African	1	1	1	1	1.5	6.75	1	0	3.077	0
rs397508609 (G126V)	0	0	0	0.000114863	African	1		1	1	1.25	8.25	5	1	1.265	0
rs1165501753 (C1410W)	0	0	0	0.000108932	East Asian	1	1	1	1	1.5	9.25	6	1	2.43	0

Table 2 Linkage disequilibrium block data for CF-related genetics from GWAS

LD block	LD SNPs > 0.8 R2	CF SNPs in LD	RegulomeDB SNPs	Missense (Bad Calls)	*Genes in LD*	*eGene (GTEx)*	Traits
chr1:205930467-205947047	16	rs4077468, rs7549173, rs117230773	rs1342063	–	*SLC26A9*	*NUCKS1*	Cystic fibrosis, type II diabetes mellitus; cystic fibrosis associated meconium ileum; Cystic fibrosis, lung disease severity measurement
chr3:195754869-195802247	21	rs3103933	–	MUC4 S585A (0)	*MUC4*	*MUC4 / LINC00969*	Cystic fibrosis, lung disease severity measurement
chr3:195802247	1	rs2688482	–	–	*MUC4*	–	Cystic fibrosis, lung disease severity measurement
chr5:33470648-33471155	1	rs139816984	–	–	–	–	cystic fibrosis associated meconium ileum
chr5:416003-591023	50	rs12188164	rs56146525, rs56279338	AHRR D627H (0)	*AHRR, EXOC3-AS1, EXOC3, SLC9A3*	*EXOC3-AS1, EXOC3, SLC9A3*	Cystic fibrosis; erythrocyte count
chr5:518319	1	rs56302516	–	–	SLC9A3	*SLC9A3*	Cystic fibrosis, lung disease severity measurement
chr5:572268-591023	26	rs57221529	rs56278696	–	–	*BRD9, AC026740.1*	Cystic fibrosis, lung disease severity measurement

Table 2 (continued)

LD block	LD SNPs > 0.8 R2	CF SNPs in LD	RegulomeDB SNPs	Missense (Bad Calls)	*Genes in LD*	*eGene (GTEx)*	Traits
chr6:32460285-32644258	227	rs9268905	rs9271589	–	*HLA-DRB5, HLA-DRB1*	*HLA-DQA2*	Cystic fibrosis; lymphoma; type I diabetes mellitus; ulcerative colitis; tonsillectomy risk measurement; antibody measurement, Epstein–Barr virus infection; inflammatory bowel disease; response to vaccine; temporal arteritis; multiple sclerosis, response to interferon beta, antibody measurement; antinuclear antibody measurement; Hodgkins lymphoma; lung carcinoma; high density lipoprotein cholesterol measurement; blood protein measurement; temporal arteritis; Epstein–Barr virus nuclear antigen 1 IgG measurement
chr7:142727839-142800839	23	rs1799886, rs3757377	–	–	*TRBV29, PRSS1, PRSS2, TRB_*	–	cystic fibrosis associated meconium ileum; susceptibility to scarlet fever measurement; alcoholic pancreatitis
chr11:34754985-34836401	180	rs7929679, rs546131, rs12793173, rs7112043	rs11032870, **rs11605381**, rs1396887	–	–	–	Cystic fibrosis; Cystic fibrosis, lung disease severity measurement; prostate specific antigen measurement, **chronic obstructive pulmonary disease**, CC16 measurement; systemic lupus erythematosus
chr13:24708681-24711776	4	rs61948108	–	–	*ATP12A*	–	cystic fibrosis associated meconium ileum
chr14:70049623-70055967	6	rs12883884	–	–	*SLC8A3*	–	Cystic fibrosis

Table 2 (continued)

LD block	LD SNPs > 0.8 R2	CF SNPs in LD	RegulomeDB SNPs	Missense (Bad Calls)	*Genes in LD*	*eGene (GTEx)*	Traits
chr16:62335996–62353040	7	rs11645366	–	–	–	–	Cystic fibrosis
chr20:50210343-50219435	12	rs2094716	rs6122889	–	*ATP9A*	–	cystic fibrosis associated meconium ileum; type II diabetes mellitus
chrX:116170939-116442508	178	rs5952223, rs7879546, rs1403543, rs3788766	rs4824377	–	*AGTR2, SLC6A14*	–	Cystic fibrosis; cystic fibrosis associated meconium ileum; Cystic fibrosis, lung disease severity measurement

motif, many known transcription factors (TFs) bound from ENCODE data, an eQTL for HLA-DQA2, and a tissue-specific enhancer of immune system cells. This LD block at chr6: 32460285-32644258 is associated with multiple immune system linked disorders (Type I Diabetes, Ulcerative Colitis, EBV infections, etc.) in addition to CF, suggesting a connection to immune response in CF patients. A total of 23 variants of the LD block have RegulomeDB scores of 2a/2b suggestive of potential to alter transcription factor binding and 14 variants a score of 3a that have a bit less evidence but potential to contribute to gene regulation.

A total of 6/15 of the LD blocks have eQTLs identified for difference genes including chr1:205930467-205947047 for NUCKS1, chr3:195754869-195802247 for MUC4/LINC00969, chr5:416003-591023 for EXOC3-AS1/EXOC3 SLC9A3, chr5:518319 for SLC9A3, chr5:572268–591023 for BRD9/AC026740.1, and chr6:32460285-32644258 for HLA-DQA2. Multiple papers have previously discussed these genes potential role in fibrosis and cystic fibrosis (Corvol et al. 2015; Wang et al. 2017, p. 3; O'Neal and Knowles 2018), with notable changes in expression for genes such as MUC4 based on CFTR (Singh et al. 2007). Out of all traits within the GWAS database, 10/15 of the LD blocks are only associated with CF based pathologies. Of the LD blocks with additional associated traits include chr5:416003-591023 with erythrocyte count, chr6: 32460285-32644258 mentioned above, chr7:142727839-142800839 with susceptibility to scarlet fever measurement/alcoholic pancreatitis, chr20:50210343-50219435 with type II diabetes mellitus, and chr11:34754985-34836401 associated with several traits including COPD.

We performed deeper analysis of this region, chr11:34754985-34836401, that overlaps CF and COPD pathology to identify novel insights on molecular genetics of this region. The region has been identified as a potential enhancer for ELF5 based on ATAC-seq, 4C-seq, and ChIP-seq (Swahn et al. 2019, p. 5), but the mechanisms of CF association for the region have not yet been resolved. The LD block is located around many genes but contains no known genes within the LD block (Fig. 4a) and no known eQTLs (Table 2). Assessment of chromosome states and known TF-binding locations in either K562 or HepG2 cells shows three potential regulatory sites within the LD block (Fig. 4b). One of the LD variants, rs11605381, falls on a PPARalpha predicted binding motif that is close to a GATA factor-binding site that could potentially regulate transcriptional activation (Fig. 4c). ChIP-Seq datasets for this region suggest potential transcription from this site including TBP and POLR2A in HepG2 cells and other factors from epithelial cells types. From the annotation data, the region of rs11605381 is an active enhancer for multiple tissues associated with CFTR expression or epithelial cells such as HepG2, rectal mucosa, lung, and intestine (Fig. 4b). The

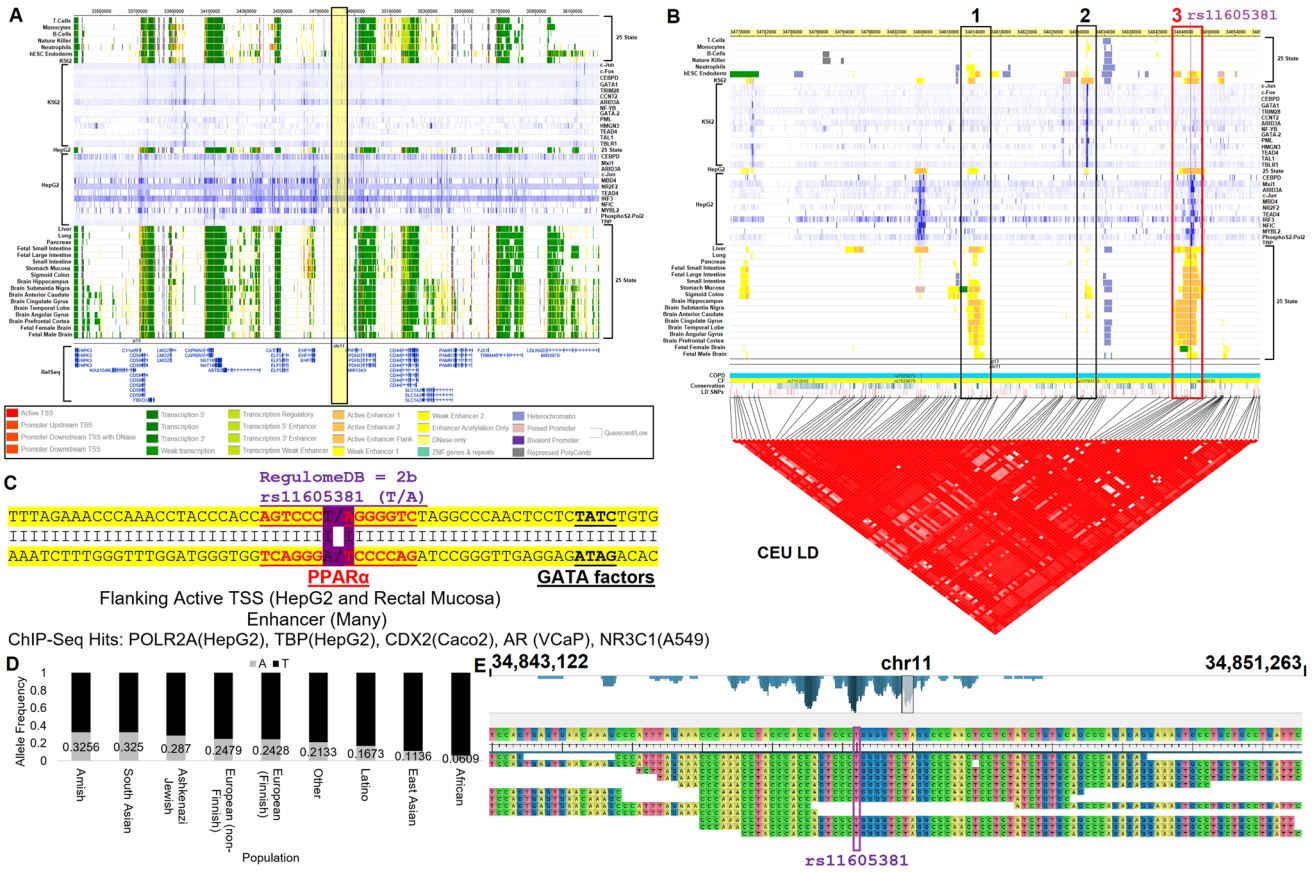


Fig. 4 CF Chromosome 11 regions that influences CF and lung pathologies. **a** Gene region around chr11:34754985-34836401 (yellow box) that associates with cystic fibrosis and chronic obstructive pulmonary disease. Data is extracted from the Roadmap Epigenomics 25-state model with the colors corresponding to the key shown below. In blue are density of ChIP-Seq-binding events from K562 and HepG2 cells. **b** Zoom in of chr11:34754985-34836401 identifying three different regulation sites within the LD block. In the red region is found the rs11605381 variant (magenta). Shown below in red is the correlation matrix of variant linkage for the CEU (Caucasian Europeans from Utah) population of the 1,000 genomes project. **c** Zoom in to sequence level for rs11605381 showing the variant near a PPARalpha potential binding site located close to a conserved GATA factor-binding site. Shown below are the known TF-binding sites from ENCODE. **d** Allele frequency for rs11605381 in different populations with A shown as gray and T in black. **e** Read mapping from Caco2 cell line RNAseq for the region surrounding rs11605381 (magenta)

rs11605381 variant is found with highest allele frequency within Amish populations (Fig. 4d). Using RNAseq datasets for epithelial cells (Caco2) we can identify reads at low depth from this region that code for noncoding RNA of unknown function (Fig. 4e). This suggests that an LD variant rs11605381 associated to both CF and COPD is found within a noncoding RNA specific to epithelial cells. It is possible this is a strong looped enhancer to ELF5 in epithelial cells that associates with weak transcription of enhancer RNA (Mikhaylichenko et al. 2018).

Defining CFTR expressing cell potential modifiers

Variants like rs11605381 linked to altered epithelial cell expression profiles suggest another strategy to map potential functional genetics of CF outside of traits captured by GWAS, where we map all genetics that can influence CFTR cells expression profiles through eQTL mapping of genes connected to CFTR expressing cells. This builds on the recent work within mouse and humans to identify the novel ionocyte cell that express CFTR as seen through single cell RNAseq (Montoro et al. 2018; Plasschaert et al. 2018). To do this we assessed CFTR expression from single cell databases (PangloaDB and the EBI Single Cell Expression Atlas) in a tissue/cell unbiased single cell scan. From 1368 single cell experiments, CFTR is found expressed in colon/intestine, pancreas, and lung datasets (Fig. 5a) particularly narrowed down from 5,586,348 cells of expression to enterocytes, cholanglocytes, ductal cells, epithelial cells, acinar cells, Paneth cells, and pulmonary alveolar type II cells (Fig. 5b). Within the lung proximal airway stromal cells, *CFTR* expression is limited to most pulmonary alveolar type II cells of human with a few luminal epithelial and basal cells (Fig. 5c). In the 53,759 single cells of 32 tissues and

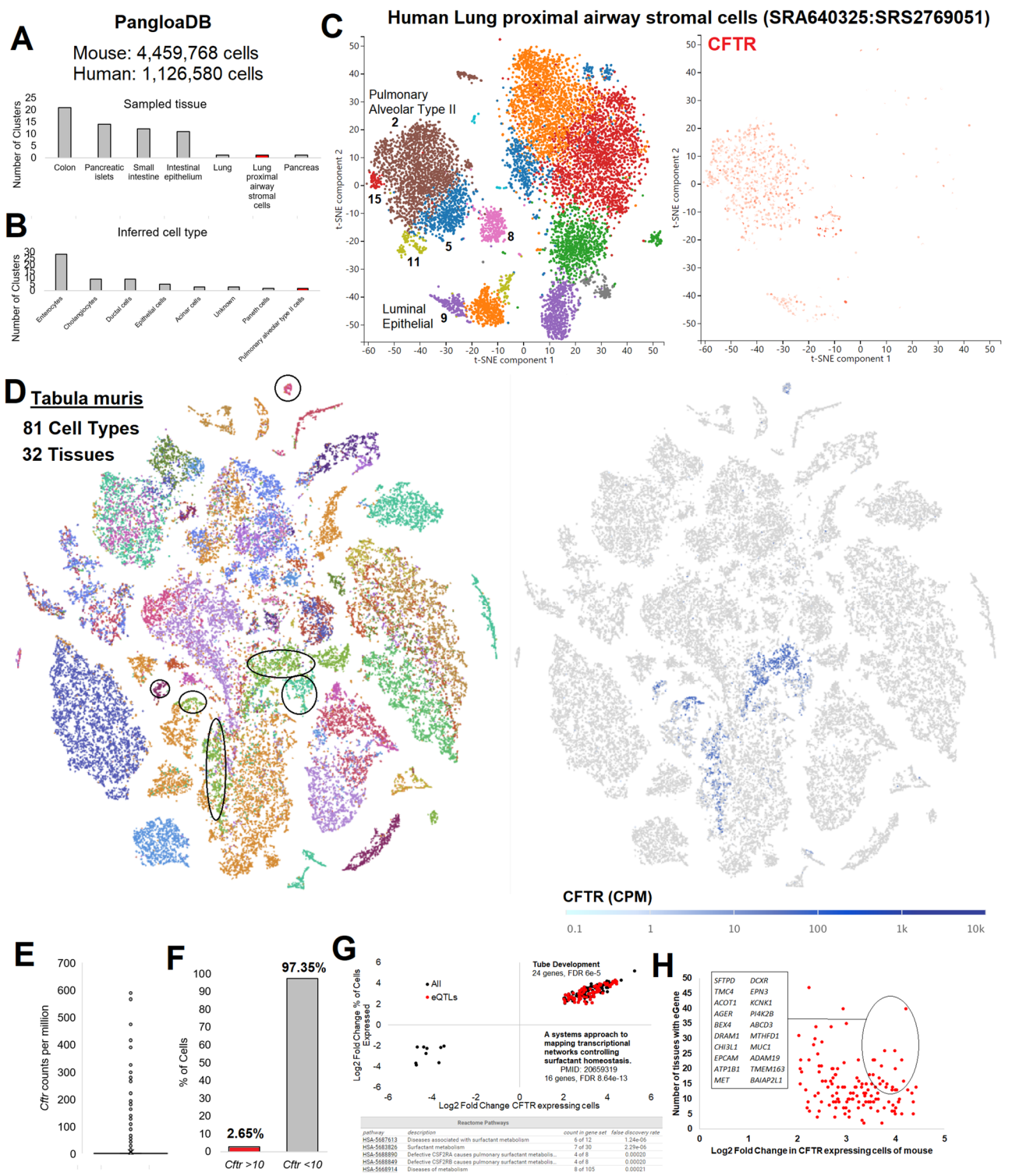
81 cell types of mouse (*Tabula muris*), *CFTR* has a very narrow expression were many of the cells cluster into 6 different groups (Fig. 5d, circled). Dissecting the lung cells for *CFTR* expression (Fig. 5e), 2.65% of cells express greater than 10 counts per million reads (Fig. 5f), allowing for segregation of gene expression in CFTR cells (Fig. 5g). Only 12 gene negatively correlate to CFTR expression while 167 positively correlate. From these genes, 24 significantly associate with tube development (FDR 6e-5) and 16 to surfactant homeostasis (Xu et al. 2010) (FDR 8.64e−13). Of the genes associated with CFTR expression, 86 have known genetics associated with changes in their expression level, known as expression quantified trait loci (eQTLs, Fig. 5g red) or expression linked genes (eGenes). Of the eGenes, multiple genes have strong enrichment in CFTR expressing cells with eQTL mapping suggesting the genes to have multiple tissue level confidence of genetics linked to expression (Fig. 5h). Five of these genes have significant human lung associated eQTLs (*ATP1B1, CHIA, SNX7, ABCD3, CHI3L1*), significantly enriching for chitin binding (FDR 1.8e−4) that has been linked to microbial responses, pediatric lung disease, and cystic fibrosis (Hector et al. 2011, p. 40; Tran et al. 2011; Mack et al. 2015; Levy et al. 2019). In addition, SFTPD has direct antimicrobial function, and has been linked with CF lung disease pathogenesis (Noah et al. 2003; Kotecha et al. 2013). This suggests the potential for variants within populations to change CFTR-based cell expression profiles and potentially cell functionality, an area of investigation that needs to be further advanced through mechanism given the small population size of GWAS.

Discussion

There has been progress towards better classifying the functional impact of CFTR variants, specifically related to CFTR modulators and potential for therapeutic intervention (Marson et al. 2016). Some evaluation of the function of missense variants has been performed with data from mutation databases such as CFTR2 (Raraigh et al. 2018). However, there are still classification difficulties, particularly for rare variants in which there are a limited number of individuals reported and limited resources for function studies. Of the characterization work of Raraigh et al. (Raraigh et al. 2018), they defined 48 variants. Yet, thousands of CFTR variants have been seen and many more will be discovered in future genomic sequencing, making it very challenging to define all variants with lab-based experiments. The presence of such undefined variants may result in an indeterminate diagnosis for some patients, loss of opportunity to benefit from CFTR modulators, and potentially unnecessary healthcare costs and uncertainty for families. Even with better classification of CFTR variants, there remains uncertainty regarding the severity of disease a patient may develop and the likelihood of developing certain symptoms (i.e. pancreatic insufficiency, pancreatitis) that may be better answered by the presence of additional genetic modifiers. Therefore, understanding the modifiers impacting disease has the potential to better guide personalized care for patients. This personalized information could also empower patients and families by providing tailored CF education and prognostic information. Thus, systematic assessment of CFTR and other genomic variants is needed for moving the CF field ahead.

GWAS have been used to investigate the genetic heterogeneity of CF, discovering loci associated with disease outcomes and severity. Several loci have been identified such as the modifier loci 11p13 and 20q13.2 (Wright et al. 2011), lung pathophysiology associated *MUC4/MUC20* on 3q29, *SLC9A3* on 5p15.3, HLA Class II on 6p21.3, *AGTR2/SLC6A14* on Xq22-q23 (Corvol et al. 2015), and *ATP12A* on chromosome 14 (Gong et al. 2019). In addition to identified modifier genes and loci that modify CF in the lungs, GWAS has also identified loci that contribute to the morbidity of other organs affected in CF (Gong et al. 2019) and CF-related diseases, such as Cystic Fibrosis-Related Diabetes (Blackman et al. 2013). Yet, from all this work, little has been done to define the causal variants and mechanisms that underlie the genetic heterogeneity. We present here that most of these genomic regions have little overlap with other traits, while one (chr6: 32460285-32644258) associates with immune system biology, and one (chr11:34754985-34836401) has overlap with COPD. The chr11:34754985-34836401 region has a gene regulation potential for a small noncoding RNA that is unique to a handful of tissues known to have CFTR expression, with a variant found near numerous transcription factor-binding sites. This finding suggests many future experiments to be performed on the noncoding RNA and how rs11605381 or other LD variants contribute to gene regulation/function. In addition, our work with eQTL mapping in CFTR expressing cells (Fig. 5) shows a promising technique to map functional biology that might contribute to genetic heterogeneity, particularly for variants that within the CF cohort have low numbers due to the limited population size. Unlike many traditional GWAS, the CF population is small in comparison to other phenotypes mapped, reducing the statistical power to overcome false discovery rate. Thus, forward thinking genetic mapping has a high probability of resolving some comodifier loci that are underpowered but suggested to overlap CFTR functional biology and cell level phenotypes when combined with the growing power of single cell transcriptomics.

Computational tools such as CADD, REVEL, SIFT, and PolyPhen2 are unable to elucidate many of the complex biological insights of variants within CFTR (Raraigh et al. 2018). While the development of tools for CFTR variant



analysis are high priority, filtering variants that are loss of function and heritable in diverse ethnicities is of critical importance. Using a population assessment of variants from gnomAD relative to the ClinVar, CFTR1, and CFTR2 databases, we have identified 18 VUS of high priority for characterization that are likely functional changes and found within diverse populations as heterozygous. Moreover, by combining a deep evolutionary analysis of CFTR in 233 species with molecular dynamics simulations of the protein within a lipid membrane we have developed a database of functional sites that can be integrated with other tools to more accurately predict loss of function variants.

◂**Fig. 5** CFTR expression and CFTR cell type eQTL mapping. **a**, **b** Expression of CFTR in the Pangloa database consisting of 4,459,768 mouse and 1,126,580 human cell expression from single cell RNAseq. Expression within different clusters of sample tissues (**a**) or inferred cell types (**b**) of the 258 total tissues and 10,399 total clusters of single cell analysis. **c** Single cell RNAseq analysis from human lung proximal airway stromal cells showing various cell clusters (left) and those cells expressing CFTR (right, red intensity corresponds to cell expression level). **d** Single cell clustering from 32 tissues and 81 cell types of mouse (left) with CFTR expression within a very limited number of cells (right, blue intensity corresponds to cell expression level). **e** The *Cftr* counts per million reads within single cells of mouse lung. **f** The percent of cells within the mouse lung that express CFTR > 10 counts per million. **g** Genes that correlate with CFTR expression in the mouse lung single cell datasets. The *x*-axis shows the Log2 fold change for each gene in cells expressing Cftr and those that do not with the *y*-axis showing the fold change in the percent of cells expressing each gene in Cftr vs non-Cftr expressing cells. Genes in red are those with known eQTLs that correlate with expression. **h** The Log2 fold change of eQTL genes in Cftr vs non-Cftr expressing cells relative to the number of tissues that the gene is known to have alterations in expression based on genetics (egene)

Amongst the most important utility of the dynamics data is the strategy to map 3D correlation of movement for all know pathogenic CFTR sites to those of VUS, suggesting functional correlations to the data correlations.

All the genomic endeavors of this project aim at one critical growing need of genomic medicine, the ability to rapidly interpret genomic variants and move them into education of clinicians and patient families. This information is important in genetic counseling and screening to see if a child will inherit CF and for proper interpretation of variants that will continue to arise de novo. Knowing the mechanism of the disease arising, plans can be made to eradicate the disease using therapeutic intervention or reproductive assistance. This gained knowledge on characteristics of different variants will be used to stratify inherited and additional de novo variants for outcomes into rapid clinical interpretation and therapy options. With knowledge of the variants, educational tools can be developed and passed along such that visual aids are available to all parties of the genomic analysis (prokoplab.com/cftr-and-cystic-fibrosis). For examples, we have developed CFTR 3D models (large: https://www.shapeways.com/product/VUMC3CJS5/cftr?optionId=144221646&li=shop-inventory; small: https://www.shapeways.com/product/BD8Z6P5NZ/cftr-small?optionId=144221588&li=shop-inventory) that highlight ΔF508 and correspond to educational material handouts also available (prokoplab.com/wp-content/uploads/2020/03/CF_Info_sheet_general_2.22.20.docx) and videos of CFTR (https://www.youtube.com/watch?v=_rVg64uTx0A&t=4s). With a continued investment into genomics and variant mechanisms, it is possible to help more CF patients while understanding additional pathologies outside of the lungs.

Availability of data and material

All material and data are presented within this manuscript.

Author contributions Performed analysis of data (MS, XL, XZ, LB, BG, KLU, DH, JWP), provided guidance on clinical variants (JMJL, JNS, SLM), oversaw project completion (CLS, MM, NL, HL, CB, JWP), wrote the manuscript (MS, JMJL, HL, CB, JWP). All authors have seen and approved the manuscript submission.

Funding This work was funded by the National Institutes of Health K01-ES025435 (JWP), Michigan State University, and Spectrum Health.

Compliance with ethical standards

Conflict of interest None of the authors have any conflicts to declare.

References

Adzhubei IA, Schmidt S, Peshkin L et al (2010) A method and server for predicting damaging missense mutations. Nat Methods 7:248–249. https://doi.org/10.1038/nmeth0410-248

Anderson MP, Gregory RJ, Thompson S et al (1991) Demonstration that CFTR is a chloride channel by alteration of its anion selectivity. Science 253:202–205. https://doi.org/10.1126/science.1712984

Apweiler R, Bairoch A, Wu CH et al (2004) UniProt: the Universal Protein knowledgebase. Nucleic Acids Res 32:D115–119. https://doi.org/10.1093/nar/gkh131

Arnold M, Raffler J, Pfeufer A et al (2015) SNiPA: an interactive, genetic variant-centered annotation browser. Bioinform Oxf Engl 31:1334–1336. https://doi.org/10.1093/bioinformatics/btu779

Blackman SM, Commander CW, Watson C et al (2013) Genetic modifiers of cystic fibrosis-related diabetes. Diabetes 62:3627–3635. https://doi.org/10.2337/db13-0510

Bobadilla JL, Macek M, Fine JP, Farrell PM (2002) Cystic fibrosis: a worldwide analysis of CFTR mutations–correlation with incidence data and application to screening. Hum Mutat 19:575–606. https://doi.org/10.1002/humu.10041

Boyle AP, Hong EL, Hariharan M et al (2012) Annotation of functional variation in personal genomes using RegulomeDB. Genome Res 22:1790–1797. https://doi.org/10.1101/gr.137323.112

Castellani C, CFTR2 team (2013) CFTR2: how will it help care? Paediatr Respir Rev 14(Suppl 1):2–5. https://doi.org/10.1016/j.prrv.2013.01.006

Cheng SH, Gregory RJ, Marshall J et al (1990) Defective intracellular transport and processing of CFTR is the molecular basis of most cystic fibrosis. Cell 63:827–834. https://doi.org/10.1016/0092-8674(90)90148-8

Choi Y, Chan AP (2015) PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinform Oxf Engl 31:2745–2747. https://doi.org/10.1093/bioinformatics/btv195

Clarke LL, Grubb BR, Gabriel SE et al (1992) Defective epithelial chloride transport in a gene-targeted mouse model of cystic fibrosis. Science 257:1125–1128. https://doi.org/10.1126/science.257.5073.1125

Corvol H, Blackman SM, Boëlle P-Y et al (2015) Genome-wide association meta-analysis identifies five modifier loci of lung disease severity in cystic fibrosis. Nat Commun 6:8382. https://doi.org/10.1038/ncomms9382

Crawford I, Maloney PC, Zeitlin PL et al (1991) Immunocytochemical localization of the cystic fibrosis gene product CFTR. Proc Natl Acad Sci USA 88:9262–9266. https://doi.org/10.1073/pnas.88.20.9262

Cutting GR, Kasch LM, Rosenstein BJ et al (1990) A cluster of cystic fibrosis mutations in the first nucleotide-binding fold of the cystic fibrosis conductance regulator protein. Nature 346:366–369. https://doi.org/10.1038/346366a0

Drumm ML, Ziady AG, Davis PB (2012) Genetic variation and clinical heterogeneity in cystic fibrosis. Annu Rev Pathol 7:267–282. https://doi.org/10.1146/annurev-pathol-011811-120900

Duan Y, Wu C, Chowdhury S et al (2003) A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. J Comput Chem 24:1999–2012. https://doi.org/10.1002/jcc.10349

ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489:57–74. https://doi.org/10.1038/nature11247

Engelhardt JF, Yankaskas JR, Ernst SA et al (1992) Submucosal glands are the predominant site of CFTR expression in the human bronchus. Nat Genet 2:240–248. https://doi.org/10.1038/ng1192-240

Farrell PM, White TB, Ren CL et al (2017) Diagnosis of cystic fibrosis: consensus guidelines from the cystic fibrosis foundation. J Pediatr 181S:S4–S15.e1. https://doi.org/10.1016/j.jpeds.2016.09.064

Feng LB, Grosse SD, Green RF et al (2018) Precision medicine in action: the impact of ivacaftor on cystic fibrosis-related hospitalizations. Health Aff Proj Hope 37:773–779. https://doi.org/10.1377/hlthaff.2017.1554

Forbes SA, Bindal N, Bamford S et al (2011) COSMIC: mining complete cancer genomes in the catalogue of somatic mutations in cancer. Nucleic Acids Res 39:D945–950. https://doi.org/10.1093/nar/gkq929

Franceschini A, Szklarczyk D, Frankild S et al (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res 41:D808–815. https://doi.org/10.1093/nar/gks1094

Franzén O, Gan L-M, Björkegren JLM (2019) PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. Database J Biol Databases Curation. https://doi.org/10.1093/database/baz046

Gong J, Wang F, Xiao B et al (2019) Genetic association and transcriptome integration identify contributing genes and tissues at cystic fibrosis modifier loci. PLoS Genet 15:e1008007. https://doi.org/10.1371/journal.pgen.1008007

GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, et al (2017) Genetic effects on gene expression across human tissues. Nature 550:204–213. https://doi.org/10.1038/nature24277

Haas BJ, Papanicolaou A, Yassour M et al (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc 8:1494–1512. https://doi.org/10.1038/nprot.2013.084

Hector A, Kormann MSD, Mack I et al (2011) The chitinase-like protein YKL-40 modulates cystic fibrosis lung disease. PLoS ONE. https://doi.org/10.1371/journal.pone.0024399

Heijerman HGM, McKone EF, Downey DG et al (2019) Efficacy and safety of the elexacaftor plus tezacaftor plus ivacaftor combination regimen in people with cystic fibrosis homozygous for the F508del mutation: a double-blind, randomised, phase 3 trial. Lancet Lond Engl 394:1940–1948. https://doi.org/10.1016/S0140-6736(19)32597-8

Kiesewetter S, Macek M, Davis C et al (1993) A mutation in CFTR produces different phenotypes depending on chromosomal background. Nat Genet 5:274–278. https://doi.org/10.1038/ng1193-274

Kotecha S, Doull I, Davies P et al (2013) Functional heterogeneity of pulmonary surfactant protein-D in cystic fibrosis. Biochim Biophys Acta 1832:2391–2400. https://doi.org/10.1016/j.bbadis.2013.10.002

Krieger E, Joo K, Lee J et al (2009) Improving physical realism, stereochemistry, and side-chain accuracy in homology modeling: four approaches that performed well in CASP8. Proteins 77(Suppl 9):114–122. https://doi.org/10.1002/prot.22570

Krieger E, Vriend G (2015) New ways to boost molecular dynamics simulations. J Comput Chem 36:996–1007. https://doi.org/10.1002/jcc.23899

Landrum MJ, Lee JM, Benson M et al (2016) ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res 44:D862–868. https://doi.org/10.1093/nar/gkv1222

Larkin MA, Blackshields G, Brown NP et al (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23:2947–2948. https://doi.org/10.1093/bioinformatics/btm404

Lek M, Karczewski KJ, Minikel EV et al (2016) Analysis of protein-coding genetic variation in 60,706 humans. Nature 536:285–291. https://doi.org/10.1038/nature19057

Levy H, Jia S, Pan A et al (2019) Identification of molecular signatures of cystic fibrosis disease status with plasma-based functional genomics. Physiol Genomics 51:27–41. https://doi.org/10.1152/physiolgenomics.00109.2018

Lonsdale J, Thomas J, Salvatore M, et al (2013) The Genotype-Tissue Expression (GTEx) project. In: Nat. Genet. https://www.nature.com/articles/ng.2653. Accessed 17 Jul 2018

MacArthur J, Bowler E, Cerezo M et al (2017) The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res 45:D896–D901. https://doi.org/10.1093/nar/gkw1133

Mack I, Hector A, Ballbach M et al (2015) The role of chitin, chitinases, and chitinase-like proteins in pediatric lung diseases. Mol Cell Pediatr. https://doi.org/10.1186/s40348-015-0014-6

Marson FAL, Bertuzzo CS, Ribeiro JD (2016) Classification of CFTR mutation classes. Lancet Respir Med 4:e37–e38. https://doi.org/10.1016/S2213-2600(16)30188-6

McKone EF, Borowitz D, Drevinek P et al (2014) Long-term safety and efficacy of ivacaftor in patients with cystic fibrosis who have the Gly551Asp-CFTR mutation: a phase 3, open-label extension study (PERSIST). Lancet Respir Med 2:902–910. https://doi.org/10.1016/S2213-2600(14)70218-8

Middleton PG, Mall MA, Dřevínek P et al (2019) Elexacaftor-tezacaftor-ivacaftor for cystic fibrosis with a single Phe508del allele. N Engl J Med 381:1809–1819. https://doi.org/10.1056/NEJMoa1908639

Mikhaylichenko O, Bondarenko V, Harnett D et al (2018) The degree of enhancer or promoter activity is reflected by the levels and directionality of eRNA transcription. Genes Dev 32:42–57. https://doi.org/10.1101/gad.308619.117

Montoro DT, Haber AL, Biton M et al (2018) A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. Nature 560:319–324. https://doi.org/10.1038/s41586-018-0393-7

Muse SV, Gaut BS (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol 11:715–724

Ng PC, Henikoff S (2003) SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res 31:3812–3814

Noah TL, Murphy PC, Alink JJ et al (2003) Bronchoalveolar lavage fluid surfactant protein-A and surfactant protein-D are inversely related to inflammation in early cystic fibrosis. Am J Respir Crit

Care Med 168:685–691. https://doi.org/10.1164/rccm.200301-005OC
Okonechnikov K, Golosova O, Fursov M, UGENE team (2012) Unipro UGENE: a unified bioinformatics toolkit. Bioinform Oxf Engl 28:1166–1167. https://doi.org/10.1093/bioinformatics/bts091
O'Neal WK, Knowles MR (2018) Cystic fibrosis disease modifiers: complex genetics defines the phenotypic diversity in a monogenic disease. Annu Rev Genomics Hum Genet 19:201–222. https://doi.org/10.1146/annurev-genom-083117-021329
Papatheodorou I, Moreno P, Manning J et al (2020) Expression Atlas update: from tissues to single cells. Nucleic Acids Res 48:D77–D83. https://doi.org/10.1093/nar/gkz947
Plasschaert LW, Žilionis R, Choo-Wing R et al (2018) A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. Nature 560:377–381. https://doi.org/10.1038/s41586-018-0394-6
Pond SLK, Frost SDW, Muse SV (2005) HyPhy: hypothesis testing using phylogenies. Bioinform Oxf Engl 21:676–679. https://doi.org/10.1093/bioinformatics/bti079
Prokop JW, Lazar J, Crapitto G et al (2017) Molecular modeling in the age of clinical genomics, the enterprise of the next generation. J Mol Model 23:75. https://doi.org/10.1007/s00894-017-3258-3
Raraigh KS, Han ST, Davis E et al (2018) Functional assays are essential for interpretation of missense variants associated with variable expressivity. Am J Hum Genet 102:1062–1077. https://doi.org/10.1016/j.ajhg.2018.04.003
Roadmap Epigenomics Consortium, Kundaje A, Meuleman W et al (2015) Integrative analysis of 111 reference human epigenomes. Nature 518:317–330. https://doi.org/10.1038/nature14248
Rogers CS, Stoltz DA, Meyerholz DK et al (2008) Disruption of the CFTR gene produces a model of cystic fibrosis in newborn pigs. Science 321:1837–1841. https://doi.org/10.1126/science.1163600
Schrijver I, Oitmaa E, Metspalu A, Gardner P (2005) Genotyping microarray for the detection of more than 200 CFTR mutations in ethnically diverse populations. J Mol Diagn JMD 7:375–387. https://doi.org/10.1016/S1525-1578(10)60567-3
Sebro R, Levy H, Schneck K et al (2012) Cystic fibrosis mutations for p. F508del compound heterozygotes predict sweat chloride levels and pancreatic sufficiency. Clin Genet 82:546–551. https://doi.org/10.1111/j.1399-0004.2011.01804.x
Singh AP, Chauhan SC, Andrianifahanana M et al (2007) MUC4 expression is regulated by cystic fibrosis transmembrane conductance regulator in pancreatic adenocarcinoma cells via transcriptional and post-translational mechanisms. Oncogene 26:30–41. https://doi.org/10.1038/sj.onc.1209764
Sun X, Yi Y, Yan Z et al (2019) In utero and postnatal VX-770 administration rescues multiorgan disease in a ferret model of cystic fibrosis. Sci Transl Med. https://doi.org/10.1126/scitranslmed.aau7531
Swahn H, Sabith Ebron J, Lamar K-M et al (2019) Coordinate regulation of ELF5 and EHF at the chr11p13 CF modifier region. J Cell Mol Med 23:7726–7740. https://doi.org/10.1111/jcmm.14646
Tabula Muris Consortium, Overall coordination, Logistical coordination, et al (2018) Single-cell transcriptomics of 20 mouse organs creates a Tabula muris. Nature 562:367–372. https://doi.org/10.1038/s41586-018-0590-4
Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10:512–526
Tamura K, Peterson D, Peterson N et al (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28:2731–2739. https://doi.org/10.1093/molbev/msr121
Tavtigian SV, Deffenbaugh AM, Yin L et al (2006) Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. J Med Genet 43:295–305. https://doi.org/10.1136/jmg.2005.033878
Taylor-Cousar JL, Mall MA, Ramsey BW et al (2019) Clinical development of triple-combination CFTR modulators for cystic fibrosis patients with one or two F508del alleles. ERJ Open Res. https://doi.org/10.1183/23120541.00082-2019
Taylor-Cousar JL, Munck A, McKone EF et al (2017) Tezacaftor-ivacaftor in patients with cystic fibrosis homozygous for Phe508del. N Engl J Med 377:2013–2023. https://doi.org/10.1056/NEJMoa1709846
Tran HT, Barnich N, Mizoguchi E (2011) Potential role of chitinases and chitin-binding proteins in host-microbial interactions during the development of intestinal inflammation. Histol Histopathol 26:1453–1464
Vecchio-Pagán B, Blackman SM, Lee M et al (2016) Deep resequencing of CFTR in 762 F508del homozygotes reveals clusters of non-coding variants associated with cystic fibrosis disease traits. Hum Genome Var 3:16038. https://doi.org/10.1038/hgv.2016.38
Wainwright CE, Elborn JS, Ramsey BW et al (2015) Lumacaftor-ivacaftor in patients with cystic fibrosis homozygous for Phe508del CFTR. N Engl J Med 373:220–231. https://doi.org/10.1056/NEJMoa1409547
Wang Y-Y, Lin Y-H, Wu Y-N et al (2017) Loss of SLC9A3 decreases CFTR protein and causes obstructed azoospermia in mice. PLoS Genet 13:e1006715. https://doi.org/10.1371/journal.pgen.1006715
Wright FA, Strug LJ, Doshi VK et al (2011) Genome-wide association and linkage identify modifier loci of lung disease severity in cystic fibrosis at 11p13 and 20q13.2. Nat Genet 43:539–546. https://doi.org/10.1038/ng.838
Xu Y, Zhang M, Wang Y et al (2010) A systems approach to mapping transcriptional networks controlling surfactant homeostasis. BMC Genomics 11:451. https://doi.org/10.1186/1471-2164-11-451
Yu H, Burton B, Huang C-J et al (2012) Ivacaftor potentiation of multiple CFTR channels with gating mutations. J Cyst Fibros Off J Eur Cyst Fibros Soc 11:237–245. https://doi.org/10.1016/j.jcf.2011.12.005
Zielenski J, Rozmahel R, Bozon D et al (1991) Genomic DNA sequence of the cystic fibrosis transmembrane conductance regulator (CFTR) gene. Genomics 10:214–228. https://doi.org/10.1016/0888-7543(91)90503-7