CFTR F508del and population structure in a cystic fibrosis population

Hanley Kingston

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

2020

Committee:

Alison Fohner

Elizabeth Blue

Program Authorized to Offer Degree:

Genetic Epidemiology

University of Washington

Abstract

CFTR F508del and population structure in a cystic fibrosis population

Hanley Kingston

Chair of the Supervisory Committee:

Alison Fohner

Department of Epidemiology

The Cystic Fibrosis Genome Project (CFGP) has assembled whole genome sequences on ~5K individuals with cystic fibrosis (CF) with the goal of identifying genetic modifiers of CF-related phenotypes. We hypothesized that the over-sampling of the clinal *CFTR* F508del haplotype in this dataset might make such studies particularly susceptible to deriving spurious associations between variants correlated with *CFTR* F508del genotype and CF-related outcomes. We assessed whether regions of the genome are associated with the *CFTR* F508del genotype by performing genome-wide association studies (GWAS's) of *CFTR* F508del genotypes and measuring the type I error rate across the genome (genomic inflation) that results when not accounting for population structure. We determined that linear mixed models with orthogonally partitioned structure (LMM-OPS) adequately controlled for the underlying relatedness and population structure within our dataset, reducing signals in genomic locations correlated with *CFTR*. Our results support that performing a GWAS of a disease-causing variant is a useful method to assess the effectiveness of principal components and genetic relatedness estimates at controlling for confounding in datasets with over-sampling of a clinal variant.

Acknowledgements

Seattle, August 20, 2020

I want to express my gratitude to my thesis chair, Alison Fohner, and my thesis advisory committee member, Liz Blue. Liz spent a huge amount of time teaching me new research tools and helping me navigate collaborations and devise research directions. Her investment in my work and education went above and beyond. Alie has offered me invaluable advice and guidance for my thesis and academic planning.

I collaborated closely on this dataset with Adrienne Stilp and William Gordon, and they helped me resolve software challenges and other errors. Adrienne wrote much of the code I used in my analysis and helped me to implement it. I also want to thank all the members of the Cystic Fibrosis Genome Project for assembling and sharing their data and ideas, and Kelsey Grinde for helping me apply findings from her dissertation to my own work.

This work is supported by the following grants from the Cystic Fibrosis Foundation: BAMSHA18XX0, CUTTIN18XX1, and KNOWLE18XX0 on behalf of the CF Genome Project.

I. INTRODUCTION

Cystic fibrosis (CF) is the most common fatal, recessive, single-gene disorder in people of European descent, affecting ~85,000 people worldwide with a global prevalence of ~1/2,500.[1] Until recently, malnutrition resulting from loss of pancreatic function frequently made CF fatal in childhood, but with recent advances in treatment, the predicted average life expectance for babies born with CF today is 41 years for women and 45 years for men in the UK, and lung disease has surpassed malnutrition as the primary cause of death.[2,3]

CF-causing variants occur in the cystic fibrosis transmembrane conductance regulator (*CFTR*) gene, which encodes an ion channel involved in regulating electrolyte levels, primarily chloride, across mucus membranes. CF-causing variants reduce membrane anion permeability and result in a buildup of mucus that obstructs the airways and pancreas, resulting in progressive respiratory illness and, with most genotypes, pancreatic insufficiency.[4] CF-causing variants have been classified into six categories, which differ in organ involvement and symptom severities. The most serious variants, Class I, cause complete loss of the protein. Class II variants are also severe, resulting in impaired traffic of the protein to the cell membrane. Class III and IV result in reduced protein function, and the least severe form of CF is caused by Class V and VI variants, which reduce the concentration and stability of *CFTR*, respectively.[1,2] For CF caused by compound heterozygosity, disease severity and pancreatic function are primarily dependent on the more functional allele. Although >1,000 variants are thought to cause CF in a recessive state, the Class II F508del variant – a deletion of 3 base-pairs (bps) that code for phenylalanine at amino acid position 508 of the protein – makes up 70% of all CF-causing variants globally, and ~82% of people with CF are carriers of at least one *CFTR* F508del variant. The *CFTR* F508del variant is most common in European populations (non-Finnish European alternate allele frequency, AAF, = 0.014) and rare elsewhere (African AAF = 0.003, East Asian AAF = 0.000).[1,5]

CFTR F508del follows a geographic cline in Europe, representing 87.2% of all CF chromosomes in a Danish sample but only 21.3% of all CF chromosomes in a Turkish sample.[6] All six CF-causing variants with global frequencies > 1% share a clinal pattern, and two of these share a similar haplotypic background (haplotype B) with *CFTR* F508del that is maintained by a 5x lower recombination rate than the genome average.[6,7] The clinal pattern of *CFTR* F508del follows known patterns of migration in Europe, and several evolutionary theories have been presented as an explanation. Some models suggest that epistatic selection with heterozygous advantage protected *CFTR* F508del carriers against diseases historically prevalent in Europe, like tuberculosis, *Vibrio cholerae*, *Salmonella typhi*, or enterotoxins.[8] Other theories suggest selection acted on *CFTR* haplotype B outside of Europe, possibly driven by diseases like cholera or typhoid fever, and this haplotype became prevalent in Europe following migration as a result of genetic drift.[7,8] Regardless of the evolutionary mechanisms by which the *CFTR* F508del frequencies were established, its clinality along known migration routes suggest that it is likely to be correlated with other variants associated with geographical space. The first two aims of our study were to 1) identify these regions, which we refer to as cross-chromosomal correlations, through a genome-wide association study (GWAS) with *CFTR* F508del genotype as the outcome, and 2) to determine whether any correlated regions have been previously identified as having long-range linkage disequilibrium (LD).

Any cross-chromosomal correlations or local structure associated with *CFTR* F508del present potential sources of confounding for genetic studies of CF-related phenotypes. The clinical presentation of CF varies widely, with phenotypes like pancreatic disease and sweat chloride levels (a common diagnostic test for CF) strongly predicted by *CFTR* causal genotype, although lung function is less so.[2] For this reason, studies of CF-related phenotypes often restrict subjects based on genotype and/or phenotype. This can introduce ascertainment bias for *CFTR* F508del, as it is both the most common CF-causing variant and a Class II variant causing a severe form of the disease. Even CF studies without genotypic or

phenotypic exclusion criteria will oversample *CFTR* haplotype B and any associated population structure as a result of its commonality among people with CF. GWAS often include principal components (PCs) as covariates to adjust for population structure within the sample to avoid such confounding. Linear mixed models with orthogonally partitioned structure (LMM-OPS) allow us to control for population structure independently of recent genetic relatedness, and our third aim was to assess whether incorporating PCs and a genetic relatedness matrix (GRM) in an LMM-OPS framework could adequately control for population structure in a GWAS of *CFTR* genotype.[9] Successful control of this population structure is essential for the success of any GWAS of CF-related traits.

Twin studies suggest that genetic modifiers outside of *CFTR* are responsible for 50% of variation in lung function and, given that 80% of people with CF die of obstructive lung disease, identifying these modifiers and those of other CF-related phenotypes, is critically important.[2] The Cystic Fibrosis Genome Project (CFGP) is a consortium between The University of Washington (UW), The University of North Carolina (UNC), and John Hopkins University (JHU) with the goal of identifying genetic modifiers of CF-related phenotypes, such as lung function, diabetes, or chronic lung infections.[10,11] In this study, we assessed whether regions of the genome are associated with the *CFTR* F508del genotype and whether we could explain away those signals with PCs. To do so, we performed GWAS's of *CFTR* F508del genotypes to assess the type I error rate across the genome (genomic inflation) that results when not accounting for population structure. Given evidence of genomic inflation, we validated the LMM-OPS framework as a method to control for the underlying relatedness and population structure within our dataset.

II. METHODS

The Cystic Fibrosis Genome Project (CFGP)

The CFGP comprises whole genome sequence (WGS) data for >5,000 CF patients across three sites: UNC, JHU, and UW. UNC contributed subjects from the Genetic Modifier Study of patients in top 25% and lowest 25% of lung disease severity and the Genetic Modifier Study of Severe Liver Disease.[12] JHU contributed subjects from the CF Twins & Sibling Study, representing 95% of known twins and sibling pairs affected by CF in the US, and the Cystic Fibrosis-Related Diabetes study. UW contributed subjects from the Early Pseudomonas Infection Control study of patients younger than 14 years of age at recruitment who tested negative for *Pseudomona arguginasa.* Clinical outcomes for all participants were collected through the national longitudinal registry, with additional information from chart review at a subset of sites.[10]

Whole Genome Sequencing

WGS data were collected for 5,199 samples by the Broad Institute Sequencing Center using their Illumina sequencing-by-synthesis protocol. Briefly, cluster amplification was performed per manufacturer's protocol by the Illumina cBot. Sequencing was performed on HiSeq X machines to produce 151bp paired-end reads. These data were processed by their Picard data-processing pipeline to generate sample-level BAM and variant call format (VCF) files aligned to the GRCh38 reference genome. Among the original 5,199 samples, 65 either failed to sequence due to low DNA yield or their sequence data did not meet basic quality control (QC) filters. A multi-sample VCF for 5,134 samples passing sequencing QC was generated using the HaplotypeCaller in the Genome Analysis Toolkit (GATK; v4.0.9.0), containing >120M single nucleotide variants (SNVs) and short insertion/deletions (indels).[13]

Sample and variant quality control

Beginning with the multi-sample VCF provided by the Broad, we applied additional sample-level and variant-level QC filters. Sample quality was assessed using evidence for contamination (Freemix estimate ≥2%),[14] high chimera rate (≥ 5%), low coverage (mean ≤29.5X, 20X coverage ≤85%, 10X coverage ≤95%), pedigree vs. empirical kinship estimates,[9,15] sample duplication or identity errors, and low support for CF diagnosis.[13] After these sample-level filters were applied, 4,966 subjects remained.[10] After restricting analysis to variants passing both the GATK Variant Quality Score Recalibration (VQSR) and hard filters (QD > 2.0; QUAL > 30.0; SOR < 3.0; FS < 60.0; MQ > 40.0; ReadPosRankSum > -8.0) with a minor allele frequency > 5% among these 4,966 samples, our analyses include 96M SNVs and indels.

CFTR genotyping

CFTR and the surrounding region (chr7:117,395,615-117,717,196 [GRCh38]) were fine-mapped in each sample to identify or verify CF-causing variants. *CFTR* genotypes in the WGS was compared with those in the national registry; registry genotypes were confirmed and discordant genotypes were resolved where possible. After expert[16] review of the *CFTR* genotypes reported in the national registry and the WGS data, causal variants were found for some of those without two known disease-causing variants where possible, [10] leaving 26 subjects excluded for unresolved C*FTR* genotypes.

Population structure and relatedness

The LMM-OPS method can be implemented through the GENESIS[17] package in R,[18] which uses the PC-Relate[9] calculation for kinship estimates and the PC-AiR[19] calculation to define PCs. Variants used to generate PCs and the GRM were limited to bi-allelic SNVs and pruned with an LD pruning threshold R^2 = 0.1 (N=178K). KING-Robust[15] was used to generate initial relatedness estimates, which were used to identify the "unrelated" set (kinship < $2^{\wedge -11/2}$) for PC-AiR. KING-Robust estimates were then adjusted for

the first 4 PCs to create a GRM using PC-Relate. Unlike PLINK and KING-Robust methods, PC-Relate kinship coefficients are independent of the ancestry proportions derived through PCs, reducing the likelihood of over-controlling for relatedness or introducing spurious associations.[9,17] The PC-AiR and PC-Relate process was repeated a second time using the results from the first round to improve precision. [15]

Statistical analyses

We define two *CFTR* F508del outcome variables which require different analytical strategies. The first is the count of *CFTR* F508del alleles observed in an individual, ranging from 0 to 2, while the second is a binary trait comparing *CFTR* F508del heterozygotes (1 *CFTR* F508del allele) to homozygotes (2 *CFTR* F508del alleles). We performed single-variant association testing using the GENESIS package:[17] we evaluated the *CFTR* F508del count outcome using a Gaussian model with inverse-transformed residuals, scaled to match the original variance of the *CFTR* count genotype, as the outcome; the *CFTR* F508del binary outcome was evaluated using a logistic model.[17] Tests were restricted to 5.7M bi-allelic SNVs with minor allele frequency (MAF) >5% and missingness < 5%. Of the 4,966 samples passing our initial QC filters, we additionally excluded one identical twin per pair (N = 27). For the binary *CFTR* F508del outcome, samples were excluded if they did not carry one or more *CFTR* F508del alleles (N = 346). For each outcome, we compared two analysis models: the baseline model adjusted for site of recruitment and the GRM, while the adjusted model also included the first four PCs as covariates. Genome-wide significance was defined using the standard threshold, 5E-08; this value is slightly above the Bonferroni-corrected p-value threshold (8.8E-09) which is overly-conservative given we have not pruned the markers for linkage disequilibrium.

We compared the PC-adjusted and unadjusted models based on their genomic inflation factors.[20] A high type I error rate, or inappropriate rejection of the null hypothesis, can occur throughout the genome as a result of confounding by population structure. Many outcomes studied using GWAS's are correlated with ancestry through the effects of geographical location on environmental exposures, sampling biases, or measurement biases. Without adjusting for differences in marker frequencies between people of different ancestries, variants correlated with ancestry and the variants in LD with them may appear to be associated with the outcome and, as a result, p-values across the genome are shifted upward by the genomic inflation factor (λ)[21] (median(test-statistic)^2/0.455). A λ value >1 indicates genomic inflation, or a median test-statistic score above expected under the chi-squared distribution, while a λ <1 indicates genomic deflation, or a median test-statistic below expected under the chi-squared distribution. Genomic inflation can lead to spurious associations, while genomic deflation is often evidence of poor study power.

III. RESULTS

Sample description

The final *CFTR* F508del count outcome analysis included 4,939 participants (1,809 JHU, 1783 UNC, 1,374 UW) and the binary outcome analysis included 4,593 participants (1,643 JHU, 1,706 UNC, 1,244 UW) (Figure 1). As shown in Table 1, the sample is multi-ethnic, but the majority of subjects (93%) reported non-Hispanic white race/ethnicity. Table 1 shows that inclusion criteria and age distribution for participants differed by study/lead institution (mean birth year = 1991 (JHU), 1982 (UNC), 2000 (UW)). The majority of participants are *CFTR* F508del homozygotes (58% of binary analysis, 63% of count analysis), and 93% of subjects in the count model are *CFTR* F508del carriers. Subjects from UNC were

more likely to be *CFTR* F508del carriers (96%) and homozygotes (75%) compared to those from the other sites. For this reason, we adjust for recruitment site in the association tests, below.

Principle component analysis

We observe that the PCs explaining the most genetic variance in our sample correspond with reported race/ethnicity values (Figure 2A&B). The parallel coordinates plot (Figure 2B) shows eigenvector values for each PC by individuals' reported race/ethnicity, illustrating that PCs 5 and up do not correspond to race/ethnicity bins and are driven by very few subjects. In the context of the reported race/ethnicity data, our sample primarily represents European ancestry, with PCs 1, 2, and 4 driven by SNVs with allele frequency differences in African, Native American, and Asian populations, respectively. PCs 3 and 6 are correlated with the *CFTR* region (Figure 3); PC 3 primarily captures variation among samples reporting white race/ethnicity but may also capture another source of variation within our dataset. The percent variance explained (Figure 4) per PC decreases drastically from PC 1 to 4. For these reasons, we used PCs 1-4 in our analysis to capture variation from population structure.

Genome-wide association between SNVs and *CFTR* F508del

A GWAS for the *CFTR* F508del count variable under the baseline model identified genome-wide significant evidence of association with variants on every chromosome tested (Figure 5A) and dramatic genomic inflation ($\lambda = 2.44$, Figure 5C). This inflation was slightly over-corrected through the inclusion of the first four PCs (Figure 5D). No variants outside of chr7 (*CFTR*) reach genome-wide significance ($p<5E-08$) in the PC-corrected model (Figure 5B).

Shared GWAS signals across baseline models

To reduce the number of significant hits and develop a clearer image of the genomic regions associated with *CFTR* F508del, we performed a GWAS limited to the binary outcome, comparing those with one vs. two copies of the *CFTR* F508del allele and found that the baseline binary model (Figure 6A) also shows high genomic inflation (λ = 1.34, Figure 6C) and shares association signals, defined by the 10 lowest p-values per analysis, in the same regions of chr2 and 12 as the baseline count model (Figure 5A). These signals fall in 7q31 (near *CFTR*), 2q21 (near *LCT* and *MCM6*), and 12p11 (near *SYT10* and *PKP2*). After adjustment for PCs 1-4 in the binary model, the association signals outside the *CFTR* region on chr7 are eliminated, with slight evidence of over-correction (λ = 0.97; Figure 6C,D); however a suggestive signal (p = 5.73E-05) remains about 13MB downstream of *LCT* and *MCM6* in the adjusted model (Figure 7B).

Genome-wide correlation between PCs and *CFTR* F508del-associated SNVs

Specific regions of the genome drive several of the early PCs, and several of these regions are associated with *CFTR* F508del. Certain regions of the genome are known to exhibit long-range LD,[22,23] and these regions of the genome are sometimes excluded when estimating PCs (Table 2) as they tend to contribute disproportionately, which can be visualized as peaks in the plot of correlation between the PCs and genomic location, (Figure 3). We identified regions of the genome with evidence of long-range LD (Supplemental Table 2), which are highlighted in green in Figure 3.[22,23] These regions intersect with some of the GWAS hits from the binary baseline model (red & purple) but show very little overlap with the GWAS hits under the binary adjusted model (blue & purple).[21] For example, PC 3 is correlated with regions on chr2, 6, and 7, containing *LCT, MHC*, (both previously documented long-range LD regions) and *CFTR*, respectively (Figure 3). *LCT* and *CFTR* both fall within association peaks in the count and binary baseline models, while *MHC* does not.

In this study, PCs 1 and 3 show the highest correlation between SNVs within 1 bp of the *CFTR* F508del (Figure 5, |R| = 0.19 & 0.12), and PCs 3, 6, 9, and 10 show spikes within the *CFTR* gene (Table 3). A preliminary analysis shows that excluding four known long-range LD regions, including the *LCT* region, suggested under the GENESIS Pipeline[24] (Table 2) from variants used in PC-generation did not reduce the correlation between PCs 3 and 6 and the *CFTR* region (|R| = 0.46 & 0.49).

IV. DISCUSSION

The extreme genomic inflation seen in the baseline (without adjusting for PCs) GWAS's of *CFTR* F508del count and binary outcomes shows that variants across the genome are associated the *CFTR*. These correlations are also evident in PC analysis, with PCs 3 and 6 showing extra correlation with *CFTR* compared to the rest of the genome and PC 3 additionally showing correlation with regions with known patterns of long-range LD, including around *LCT*. Two long-range LD regions, 2q21 and 12p11, are also associated with *CFTR* F508del in our baseline model GWAS's, but these cross-chromosomal correlations can be largely accounted for by PCs 1-4, which suppress the association between these regions and *CFTR* F508del outcomes. This means that PC-adjustment in the LMM-OPS framework can address these correlations and is likely also an appropriate method to prevent confounding by *CFTR* genotype in studies of CF modifiers within the CFGP.

Association testing of the *CFTR* F508del outcome reveals that cross-chromosomal correlations with *CFTR* include several regions linked to population structure and selection and that PCs 1-4 adequately control for this correlation. Of the 11 peaks that reached genome-wide significance in the binary *CFTR* F508del association tests under the baseline model, we focused on two that were also in the top five association signals in the baseline count model and that fell within regions of the genome previously identified as having long-range LD: rs533344 in 2q21 and rs949473 in 12p11 (Supplemental Table 3).[22] rs533344 falls

within 750KB of *LCT* and *MCM6* (a regulator of *LCT*), genes in which variation follows geographical clines as a result of selection.[25] The frequency of the *MCM6* alleles for higher lactase-persistence is higher in populations with European ancestry and lower in those with Asian and Southern African ancestry and, within Europe, follows a similar relative distribution to *CFTR* F508del.[1,25] The peak on 12p11 encompasses only one gene, *SYT10*, which has been associated with selection in humans[26] and shows modest evidence of selective constraint (loss-of-function observed/expected upper bound fraction = 0.75).[5] These results suggest that we should be suspicious of associations found between CF modifiers and genomic regions with histories of selection, like *LCT*/*MCM6* or *SYT10*; however, the reduction of the associated peaks with the inclusion of the first four PCs in the binary model suggests that PC adjustment adequately captures the cross-chromosomal correlation between *CFTR* and such long-range LD regions. The peak from rs1911632, about 13MB downstream of *LCT* and *MCM6*, does retain a suggestive (but not genome-wide significant) p-value in the association tests for the binary *CFTR* F508del model with PCs, and whether this peak is driven by *LCT* or is independently correlated with *CFTR* requires further investigation.

Although including PCs as covariates in GWAS's can control for population structure in the test, the selection of PCs must be done with care. Figure 3 shows that, among variants used in the association testing, PC 3 is most highly correlated with variants in the *LCT* and *CFTR* regions, both of which represent peaks in the binary baseline model. Controlling for PCs driven by local structure, as indicated by PC-correlation plots, can result in over-correction and a decrease in the power to detect true causal associations with this region. However, recent research suggests that controlling for PCs driven by two or more distinct sites of local ancestry, where one of those sites is truly associated with the outcome and other(s) is neutral, can actually induce spurious association between the truly neutral locus and the outcome as a result of collider bias.[22,23] Such a finding raises concern that controlling for PC 3 could induce a spurious association between loci near *LCT* or *MHC* with the *CFTR* genotype outcome (or in a

GWAS of a CF modifier), as these loci contribute disproportionately to PC 3 eigenvector values alongside *CFTR*. However, we see the opposite effect – the signal for *LCT* is strongest in the baseline model and significantly reduced through the inclusion of PCs 1-4 (Figure 6 & 7). The lack of an association signal in the *MHC* region in the baseline or adjusted model also supports that including PC 3 does not induce spurious associations. It should be noted that PC 6 is also correlated with *CFTR* (Figure 3) but lacks any other genomic regions with strong correlation peaks and, therefore, cannot not tag any other hits in our association results. PC 6 cannot act as a collider and, as it is driven by one region of local structure, is appropriate to exclude. Most evidence for collider bias in GWAS's come from studies of admixed populations, and more research is needed to determine whether collider bias plays a role in studies with few admixed individuals, like this one.

Our study has several limitations. We did not evaluate the utility of other frameworks for PC and GRM generation in this dataset and cannot say whether other methods of PC and GRM development are adequate to control for population structure in studies where the outcome is caused by a clinal variant. It is possible that some of the genomic inflation within the baseline count model resulted because *CFTR* F508del genotype is not a continuous, normally-distributed trait (Figure 1), an assumption of the Gaussian model. However, the genomic inflation in this model, compared to the binary model, may also reflect greater genetic heterogeneity from including homozygotes for non-F508del *CFTR* alleles, as they are more likely than *CFTR* F508del carriers to have non-European ancestry. Modeling the *CFTR* F508del count using a Poisson distribution may better capture the distribution of the outcome. An additional limitation in our study is that we did not capture the full diversity, in terms of haplotypic background and phenotypic effect, within non-F508del *CFTR* alleles. For example *CFTR* F508del heterozygotes with a non-functional second allele are likely to be more phenotypically similar to *CFTR* F508del homozygotes than to heterozygotes with one partially-functioning allele.[2]

Although treatments for CF have advanced rapidly, it is still considered a fatal disease, and patients require medical interventions throughout their lives.[1] For most people with CF, the causal variants are known, but there is a great need to understand genetic modifiers of CF outcomes. Combining subject data from multiple studies can greatly increase the power to detect these genetic modifiers, but the ample close genetic relatedness in CFGP, which includes participants obtained as part of a twins and siblings study; the combining of multiple studies with different genotypic and phenotypic inclusion criteria; and the clinality of the *CFTR* F508del variant necessitate careful control of population structure and relatedness. Methods to control for population structure in such a dataset must be robust not only to typical sources of population structure but also to population structure signals magnified by ascertainment bias. Our results support that performing a GWAS of the disease-causing variant is a useful method to assess the effectiveness of PCs at controlling for confounding by regions of the genome correlated with a disease-causing variant with a history of selection. We show that in this dataset, SNVs across the genome are associated with *CFTR* F508del but that accounting for both pedigree- and population-based genetic correlations between subjects can appropriately control for this population structure, avoiding inflated test statistics and spurious associations in a GWAS of *CFTR* F508del genotype.

References

1. Lopes-Pacheco Miquéias. CFTR Modulators: The Changing Face of Cystic Fibrosis in the Era of Precision Medicine. *Front Pharmacol.* 2020;10.

2. Cutting Garry R. Cystic fibrosis genetics: from molecular understanding to clinical application. *Nat Rev Genet.* 2015;16(1).

3. Keogh Ruth H., Szczesniak Rhonda, Taylor-Robinson David, Bilton Diana. Up-to-date and projected estimates of survival for people with cystic fibrosis using baseline characteristics: A longitudinal study using UK patient registry data. *J Cyst Fibros.* 2018;17(2).

4. Johns J. D., Rowe Steven M. The effect of CFTR modulators on a cystic fibrosis patient presenting with recurrent pancreatitis in the absence of respiratory symptoms: a case report. *BMC Gastroenterol.* 2019;19(1).

5. Karczewski Konrad J., Francioli Laurent C*., et. al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature.* 2020;581(7809).

6. Mateu Eva, Calafell Francesc*, et. al.* Can a place of origin of the main cystic fibrosis mutations be identified? *Am J Hum Genet.* 2002;70(1).

7. Gradient of distribution in Europe of the major CF mutation and of its associated haplotype. European Working Group on CF Genetics (EWGCFG). *Hum Genet.* 1990;85(4).

8. Wiuf C. Do delta F508 heterozygotes have a selective advantage? *Genet Res.* 2001;78(1).

9. Conomos Matthew P, Reiner Alexander P, Weir Bruce S, Thornton Timothy A. Model-free Estimation of Recent Genetic Relatedness. *The American Journal of Human Genetics.* 2016;98(1).

10. Aksit,Melis A. The Cystic Fibrosis Genome Project: A Multi-Phenotype Whole Genome Sequencing Resource [unpublished report]. Johns Hopkins University Department of Genetic Medicine2020.

11. Ling et. al, 2020, The Cystic Fibrosis Genome Project: A Multi-Phenotype Whole Genome Sequencing Resource [abstract].

12. Stonebraker Jaclyn R., Ooi Chee Y*., et. al.* Features of Severe Liver Disease with Portal Hypertension in Patients With Cystic Fibrosis. *Clin Gastroenterol Hepatol.* 2016;14(8).

13. Whole Genome Sequencing [cited Aug 15, 2020]. Broad Institute: Genomic Services. Available from: http://genomics.broadinstitute.org/products/whole-genome-sequencing.

14. Jun Goo, Flickinger Matthew*, et. al.* Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am J Hum Genet.* 2012;91(5).

15. Manichaikul Ani, Mychaleckyj Josyf C*., et. al.* Robust relationship inference in genome-wide association studies. *Bioinformatics.* 2010;26(22).

16. McCague Allison F., Raraigh Karen S.*, et. al.* Correlating Cystic Fibrosis Transmembrane Conductance Regulator Function with Clinical Features to Inform Precision Treatment of Cystic Fibrosis. *Am J Respir Crit Care Med.* 2019;199(9).

17. Gogarten S. M., Sofer T.*, et. al.* Genetic association testing using the GENESIS R/Bioconductor package. *Bioinformatics.* 2019.

18. R Core Team (2019). V3.6.1. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

19. Conomos Matthew P., Miller Michael B., Thornton Timothy A. Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. *Genet Epidemiol.* 2015;39(4).

20. Devlin B., Roeder Kathryn. Genomic Control for Association Studies. *Biometrics.* 1999;55(4).

21. Price Alkes L., Patterson Nick J.*, et. al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics.* 2006;38(8).

22. Price Alkes L., Weale Michael E.*, et. al.* Long-Range LD Can Confound Genome Scans in Admixed Populations. *Am J Hum Genet.* 2008;83(1).

23. Grinde Kelsey. Statistical Inference in Admixed Populations [dissertation]. University of Washington; 2019.

24. TOPMed analysis pipeline: UW Genetic Analysis Center; 2017 [updated Sep 7; cited July 20, 2020]. Available from: github.com/UW-GAC/analysis_pipeline.

25. Itan Yuval, Jones Bryony L.*, et. al.* A worldwide correlation of lactase persistence phenotype and genotypes. *BMC Evol Biol.* 2010;10.

26. Simonson Tatum. Genetic Adaptation to High Altitude in Tibetans [dissertation]. University of Utah; 2011.

Figures

Figure 1. Count of *CFTR* F508del alleles per subject, stratified by study site. Definitions: JHU = Johns Hopkins University, UNC = University of North Carolina, UW = University of Washington.

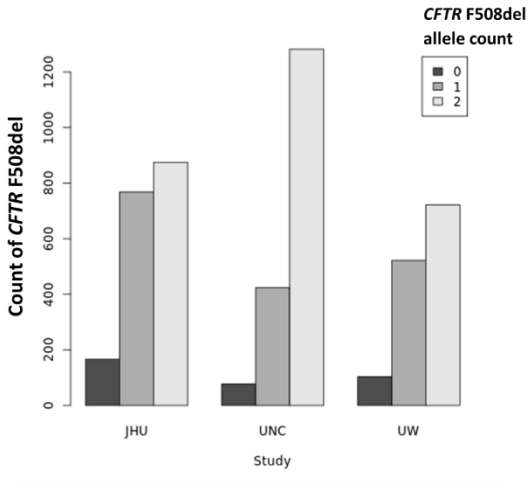


Figure 2: Population structure within the CFPG as measured by principle components. Reported race and ethnicity were pulled from the CF registry. (A) Pairwise principle component (PC) plots for PCs 1-6. (B) Eigenvector values for PCs 1-18.

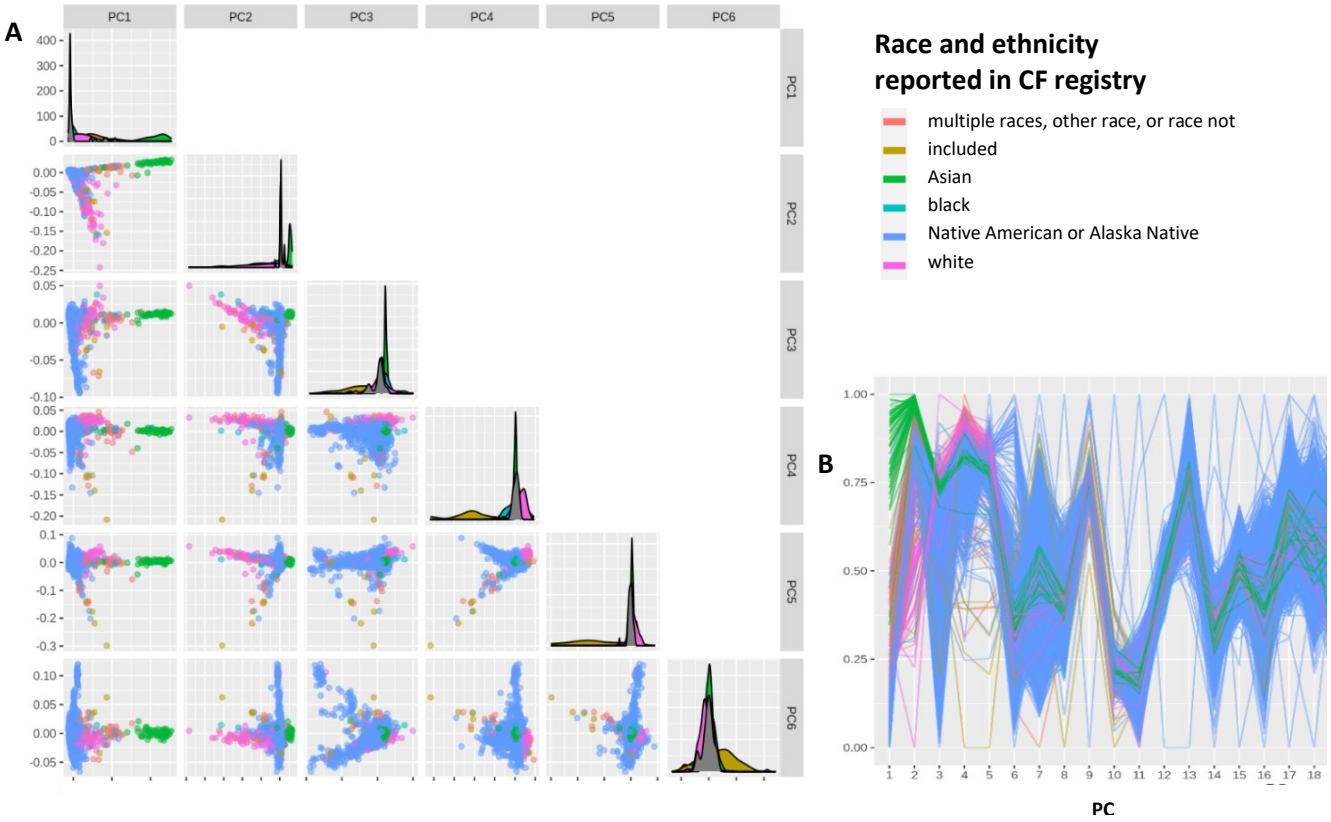


Figure 3: Correlation between principle components (PCs) 1-10 and genomic position. Highlights show the 10 loci with the lowest p-value peaks from the binary baseline model (red) and binary adjusted model (blue) with peaks shared by both models (purple). Regions with prior evidence of long-range LD are green.[22,23] Variant positions are on the GRCh38 map.

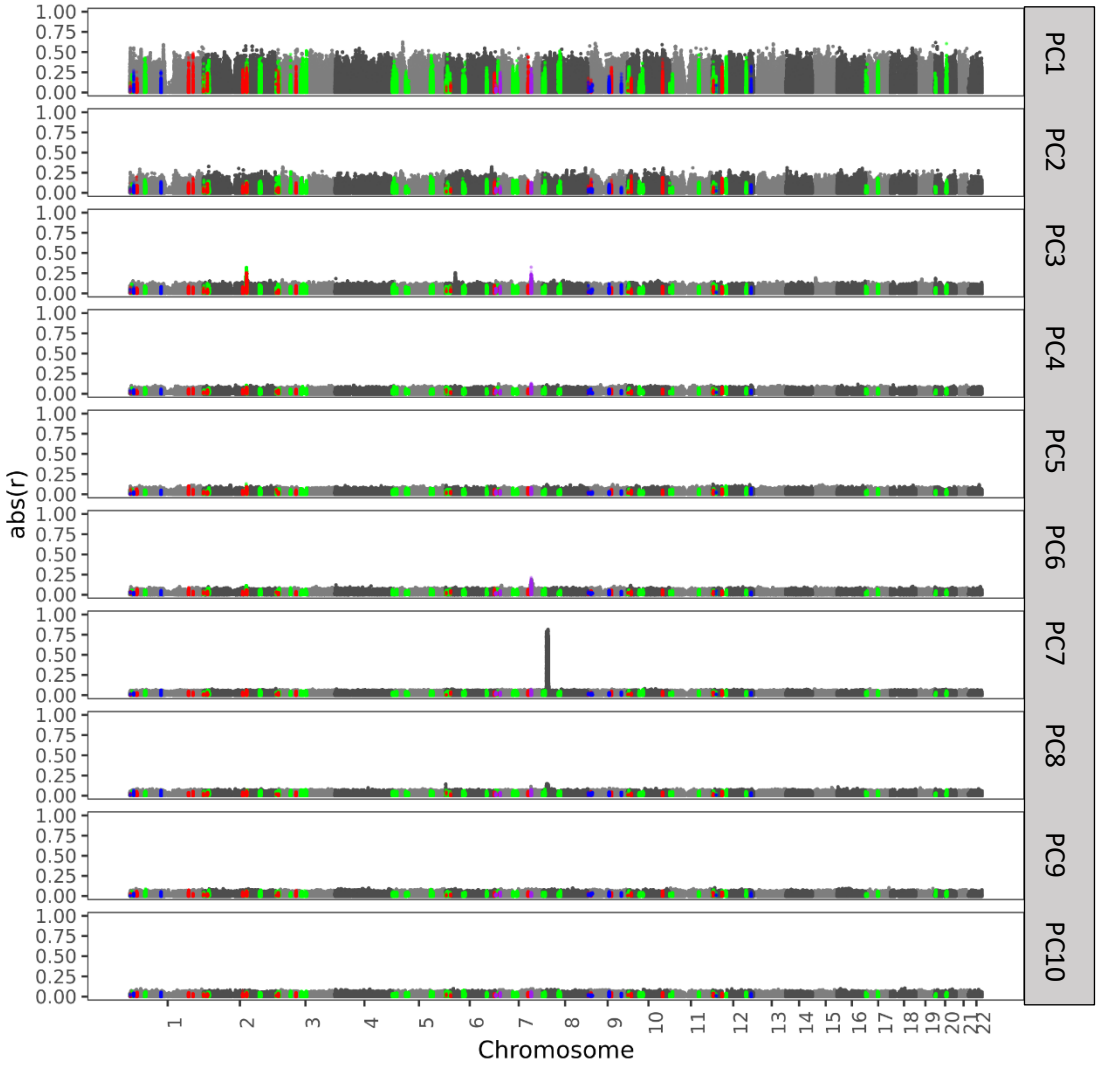


Figure 4. Scree plot of percent of the genetic variance explained by PCs 1-18.

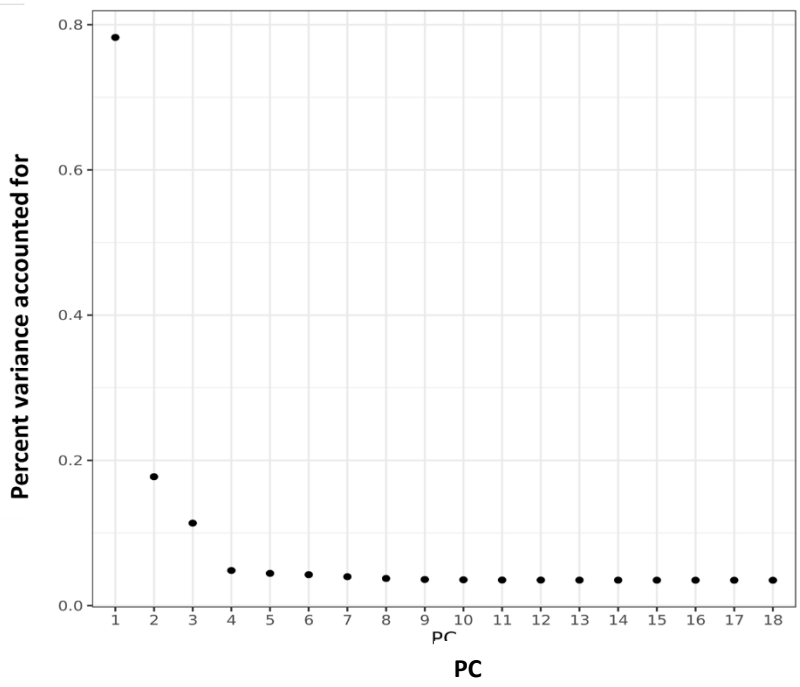


Figure 5: Genome-wide association studies for the count of *CFTR* F508del alleles (one, two, or three alleles), adjusted for site and relatedness. Panels A and C: Results from the baseline model, panels B and D: results from PC-adjusted model. Panels A and B illustrate association signals, measured as -$\log_{10}$(p-values), truncated at p = 1E-25 (A) and p = 1E-12 (B), as the peak at *CFTR* on chr7 reaches p <1E-300 under both models. Significance levels: 5E-08, 5E-09, and 1E-09 are indicated by horizontal dashed lines. Panels C and D are quantile-quantile plots comparing observed vs. expected values under the beta-distribution, with lines at the equivalence and the 95% CI under the beta distribution.

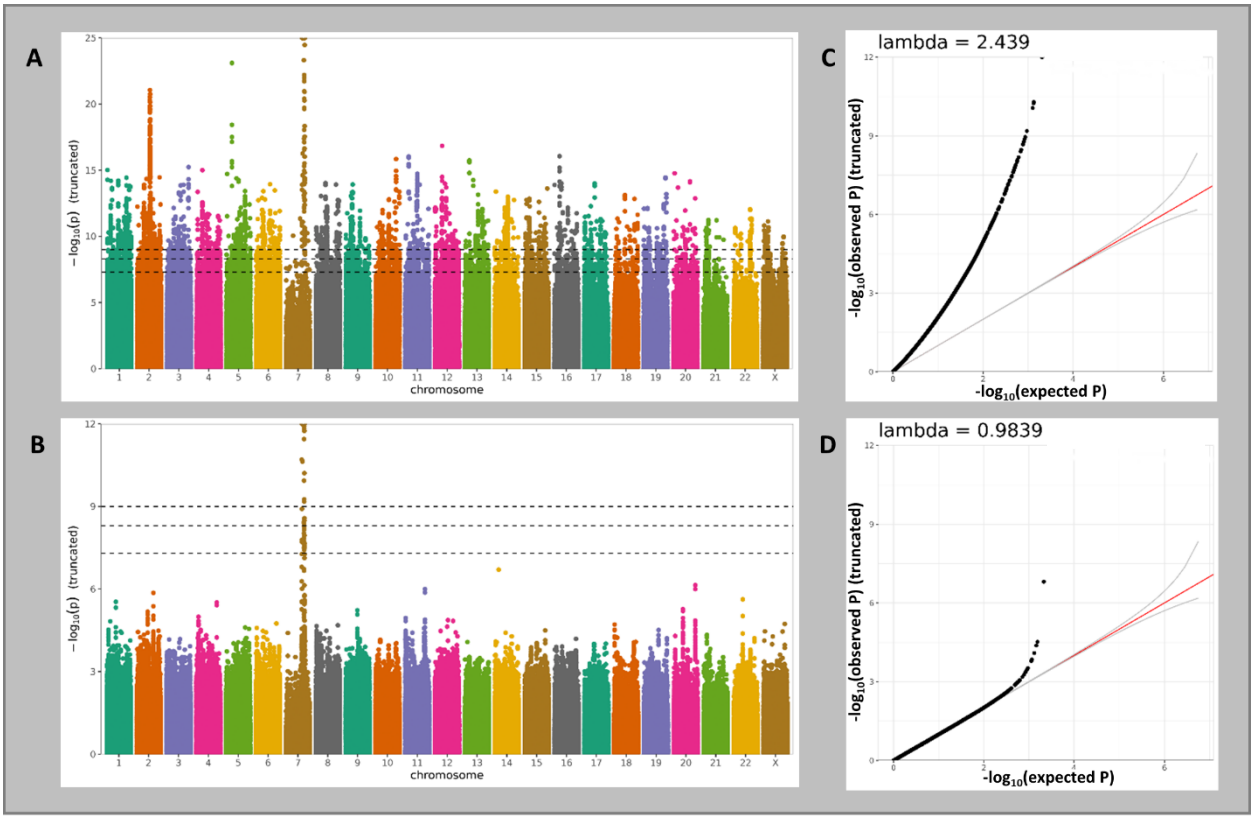


Figure 6: Genome-wide association studies for *CFTR* F508del homozygotes/heterozygotes (one or two alleles), adjusted for site and relatedness. Panels A and C: Results from the baseline model, panels B and D: results from PC-adjusted model. Panels A and B illustrate association signals, measured as -$\log_{10}$(p-values), truncated at p = 1E-12, as the peak at *CFTR* on chr7 reaches p <1E-300 under both models. Significance levels: 5E-08, 5E-09, and 1E-09 are indicated by horizontal dashed lines. Panels C and D are quantile-quantile plots comparing observed vs. expected values under the beta-distribution, with lines at the equivalence and the 95% CI under the beta distribution.

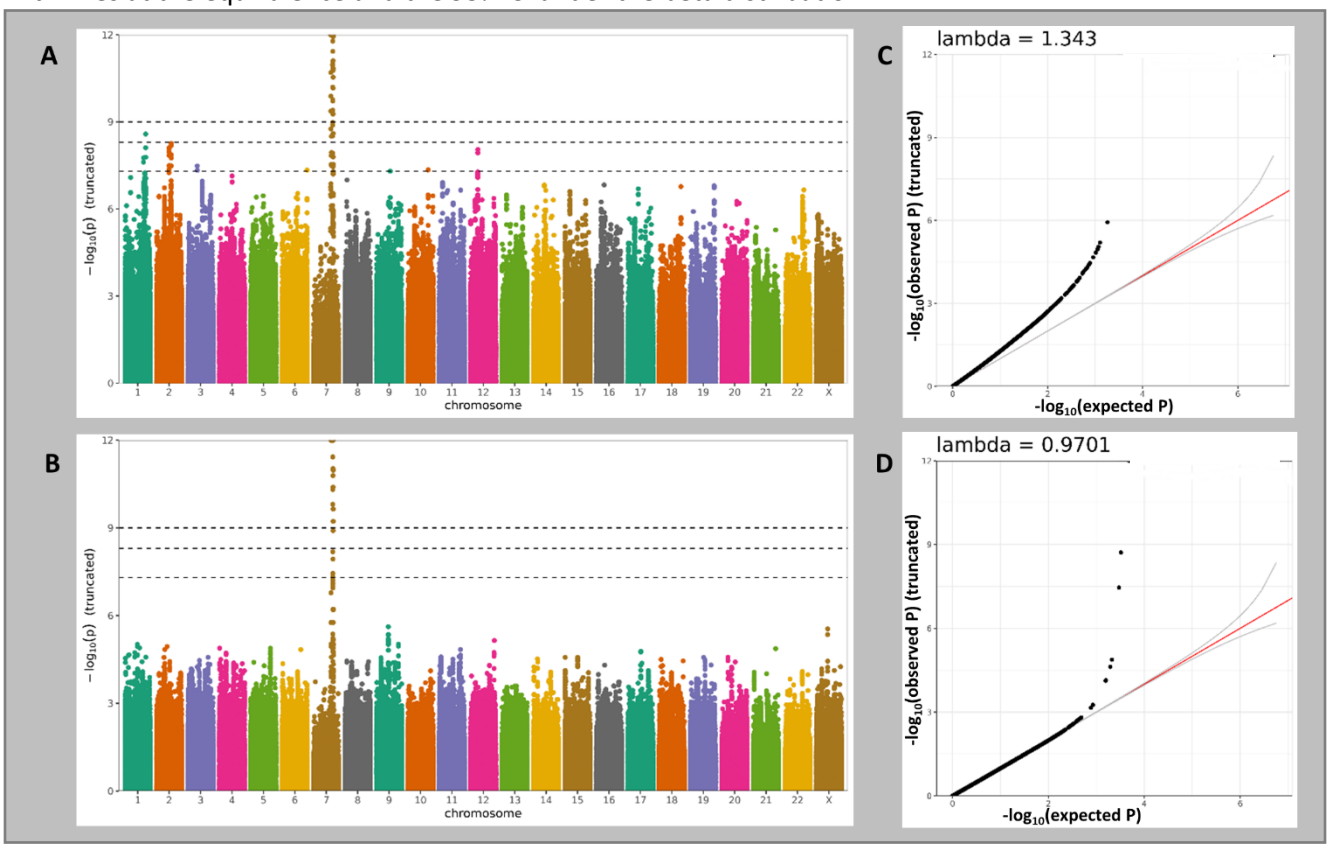
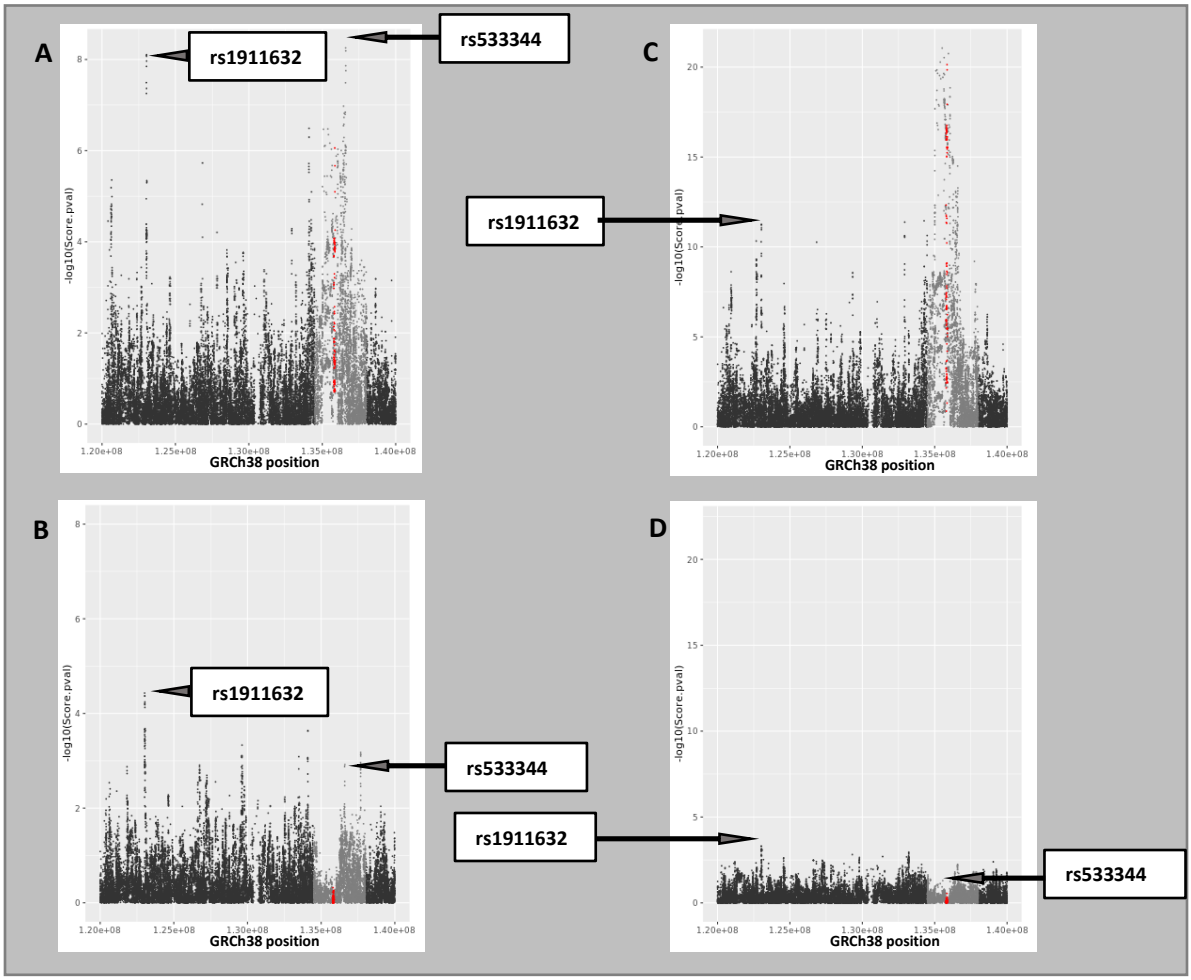


Figure 7: Evidence of association between *CFTR* F508del and the *LCT* locus without (A,C) and with (B,D) adjustment for principal components. Panels A,B are the results under the binary *CFTR* F508del outcome and panels C and D are the results for the count of *CFTR* F508del alleles. The *LCT* and *MCM6* genes are highlighted in red, while the light grey region depicts the surrounding region (chr2, 134.5–138 Mb) with prior evidence of long-range LD.[22,23] The peaks (rs1911632 and rs533344) from Supplemental Table 1 are labeled.



Tables

Table 1: Characteristics of subjects included in analysis. Binary model in parenthesis.

Study	Twins and Sibs (TSS)	Cystic Fibrosis-Related Diabetes (CFRD)	Genetic Modifier Study	Genetic Modifier Study of Severe Liver Disease Study	Early Pseudomonas Infection Control (EPIC)	Total
lead Institution	John Hopkins		University of North Carolina		University of Washington	
design (inclusion criteria)	Family and population-based (twin and sibling sets concordant for CF)		Extremes of phenotype (top of bottom 25% of SaKnorma; pancreatic insufficient CFTR variant; age > 7)		Longitudinal cohort (age < 14 when enrolled; neg. for *Pseudomona arguginasa*)	
mean birth year (range)	1991 (1943-2011)		1982 (1946-2007)		2000 (1992 – 2006)	1900 (1943-2011)
mean age diagnosed, years	2.4		2.4		0.9	2.0
percent *CFTR* F508del carriers	91		96		92	93
percent *CFTR* F508del homozygotes	48 (53)		72 (75)		54 (58)	58 (63)
percent male	52		56		50	53
race/ethnicity						
Asian*	11 (8)		0 (0)		2 (2)	13 (10)
black*	31 (19)		25 (19)		35 (28)	91 (66)
Hispanic*	74 (57)		42 (22)		42 (36)	138 (115)
Native American / Alaska Native*	10 (10)		5 (5)		10 (10)	25 (25)
white*	1647 (1515)		1702 (1637)		1230 (1148)	4579 (4300)
Total	1809 (1643)		1783 (1706)		1347 | 1244	4939 (4593)

*based on CFGP reports

Table 2: Evidence of association between *CFTR* F508del and variants within regions with long-range linkage disequilibrium (LD). Regions with patterns of long-range LD are defined as those suggested for exclusion from PC estimation under GENESIS Pipeline.[17] The association results are for the binary *CFTR* F508del outcome, with the variant with the smallest p-value within the region reported. Notes: genes or structural variants of interest within the region. Positions are given relative to the GRCh38 reference genome.

Position	**PC-adjusted model**		**Baseline model**		**Notes**
	rsID	**p-value**	**rsID**	**p-value**	
2 :129125957-139525961	rs995642	2.31E-04	rs533344	5.58E-09	*LCT*
6:24091793-38924246	rs2766538	1.44E-04	rs377743	1.41E-06	Encompasses *MHC*
8: 675507-13598120	rs6997954	3.57E-04	rs1816014	9.99E-08	Inversion
17 :42394456-46567318	rs8068816	1.28E-04	rs62066715	2.06E-04	Inversion

Table 3: Maximum correlation between variants in *CFTR* and each of the first 10 principal components. Variants are not restricted to those included in association testing. PC: principal component. *CFTR* region defined by its position on the GRCh38 reference (7: 117,480,025-117,668,665).

PC	1	2	3	4	5	6	7	8	9	10
max|R|	0.51	0.28	0.46	0.21	0.30	0.51	0.07	0.15	0.49	0.44

Supplemental Tables

Supplemental Table 1: Regions of the genome with significant evidence of association with the binary *CFTR* F508del outcome under the baseline model. Significant p-value threshold: 5E-08. Association model compares those with 1 vs. 2 copies of *CFTR* F508del and adjusts for site and a genetic relatedness matrix. Sequence positions of association peaks are provided on the GRCh38 map. * other genes in region are not listed. ** within 2MB of centromere.

lead SNV	rsID	AAF (non-Finnish Eur | all)[5]	MAF in sample	p-value	Association peak	genes in range	next nearest gene(s)
7:117523562	rs7802924	0.09 | 0.07	0.15	<1E-300	7:116650000-118000000	*CFTR**	
1:213884381	rs853741	0.05 | 0.25	0.06	2.62E-09	1:213750000-214050000	*PROX1*	*AC096639; LINC00538*
2:136591417	rs533344	0.28 | 0.45	0.30	5.58E-09	2:133000000-139000000	*ZRANB3; RAB3GAP1; R3HDM1; ACMSD; LCT; MCM6; TMEM163; MGAT5; DARS CXR4; THSD7B; NCKAP5*	
2:123034741	rs1911632	0.18 | 0.32	0.20	7.84E-09	2:123000000-123250000	*AC062020; LINC01826*	
12:33254158	rs949473	0.14 | 0.33	0.16	8.99E-09	12:33000000-33445000**	*SYT10*	
7:104037729	rs66918163	0.04 | 0.27	0.05	1.03E-08	7:103700000-104100000	*RELN*	*ORC5; LHFPL3*
1:198452278	rs2813164	0.69 | 047	0.33	1.74E-08	1:197700000-198800000	*AL450352; ATP6V1G3; NEK7; PTPRC; LHX9; C1orf53; DENND1B*	*ATP6V1G3*
3:67733121	rs11127729	0.04 | 0.16	0.05	3.28E-08	3:67350000-67800000	*SUCLG2*	*TFA1; KBTBD8*
10:111762996	rs1923653	0.06 | 0.22	0.10	4.47E-08	10:111500000-112000000		*GPAM; TECTB; ACSL5; ZDHHC6; ADRA2A; VTI1A; SHOC2*
6:168182326	rs9455973	0.09 | 0.22	0.10	4.56E-08	6:168100000-168300000	*LOC105378137; AL606970; LOC101929420; DACT2; FRMD1*	
9:84291192	rs6559779	0.09 | 0.25	0.19	4.95E-08	9:84050000-84550000	*SLC28A3*	*NTRK2, RMI1*

Supplemental Table 2: Regions of the genome with suggestive evidence of association with the binary *CFTR* F508del outcome under the adjusted model. Suggestive p-value threshold = 1E-06. Association model compares those with 1 vs. 2 copies of *CFTR* F508del and adjusts for site, principal components (PCs) 1-4, and a genetic relatedness matrix. Sequence positions are given on the GRCh38 map. * other genes in region are not listed. ** within 2MB of centromere.

SNP	rsID	AAF (non-Finnish Eur | all)[5]	MAF in sample	p-value	chr:start-end hg38	genes in range	next nearest gene(s)
7:117523620	rs7786196	0.30 | 0.25	0.23	<1E-300	7:116650000-118000000	*CFTR**	
9:76348207	rs9987416	0.32 | 0.37	0.39	2.41E-06	9:76250000-76450000	*PCSK5; RFK; GCNT1*	
X:64815513	rs146563636	0.06 | 0.04	0.05	2.86E-06	X:63200000-65500000**	*AMER1; MTMR8; ARHGEF9; ZC4H2; SPIN4; ZC3H12B*	*LASIL; MSN; VSIG4; ASB12; HEPH; EDA2R*
7:124136003	rs56176926	0.17 | 0.16	0.20	4.80E-06	7:123850000-124650000	*AC006148*	*TMEM229A; SPAM1*
12:127389705	rs11059101	0.13 | 0.15	0.14	7.09E-06	12:12735000-12740000	*APOLD1*	*CDKN1B*
9:118964565	rs10984344	0.10 | 0.09	0.10	9.47E-06	9:118900000-119050000		*BRINP1; LINC02578*
1:107543392	rs12042180	0.12 | 0.10	0.12	9.64E-06	1:107300000-107800000		*VAV3; NTNG1*

Supplemental Table 3: Evidence for association between binary *CFTR* F508del outcome and variants within regions reported to have long-range LD in those with European ancestry, adjusted for site and relatedness. Relatedness adjustment is through a genetic relatedness matrix adjustment. Position is base-pair position on chromosome 7 using human genome reference build GRCh38.

LD block [22,23]	lowest p-value from baseline model		lowest p-value from adjusted model	
	rsID	p-value	rsID	p-value
1:48000000-52000000	rs2244156	4.84E-05	rs3827730	2.98E-03
2:86000000-100500000	rs79964908	7.70E-05	rs79964908	1.15E-05
2:134500000-138000000	rs533344	5.58E-09	rs7583594	6.67E-04
2:183000000-190000000	rs1850277 & rs1850275	7.37E-05	rs1850277 & rs1850275	1.52E-03
3:47500000-50000000	rs7434107	9.73E-05	rs34711187	2.35E-03
3:83500000-87000000	rs62264006	3.27E-05	rs12106798	3.44E-05
3:89000000-97500000	rs28604999	6.49E-06	rs72914997	3.02E-03
5:44500000-50500000	rs1482698	4.13E-02	rs1263948864	6.75E-02
5:98000000-100500000	rs7443143	6.50E-07	rs7443143	5.20E-05
5:129000000-132000000	rs10069257	5.16E-05	rs13160692	1.57E-03
5:135500000-138500000	rs11746304	4.62E-04	rs6864240	6.77E-04
6:25500000-33500000	rs377743	1.41E-06	rs1796518	1.59E-03
6:57000000-64000000	rs1325665473	1.60E-03	rs1325665473	3.59E-04
6:140000000-142500000	rs6570439	1.58E-03	rs80011279 & rs78676868	3.34E-03 &
7:55000000-66000000	rs10274542	4.07E-05	rs1392975109	2.26E-04
8:8000000-12000000	rs6601549	1.13E-04	rs6992153	6.91E-04
8:43000000-50000000	rs4526375	5.42E-05	rs7003908	7.28E-04
8:112000000-115000000	rs10095591	9.89E-06	rs73357266	4.45E-04
10:37000000-43000000	rs74844346	1.72E-02	rs72777186	1.01E-02
11:87500000-90500000	rs12272487	1.89E-04	rs10831462 & rs10831463	1.52E-03
11:46000000-57000000	rs11512800	6.32E-05	rs11512800	6.22E-05
12:33000000-40000000	rs949473	8.99E-09	rs7975300	2.80E-04
12:109500000-112000000	rs7300252	3.22E-04	rs12301787	2.65E-03
17: 40000000-43000000	rs803127	2.96E-05	rs2290065	1.17E-03
20:32000000-34500000	rs138420028	9.89E-06	rs138420028	3.86E-05