Title: GENETIC MODIFIERS OF CYSTIC FIBROSIS

Abstract: Disclosed herein are compositions and methods for and treating Cystic Fibrosis lung disease severity and/or secondary manifestations, including meconium ileus and CF related liver disease.
Genetic Modifiers of Cystic Fibrosis

Related Applications
This application claims the benefit of U.S. Provisional Application Serial No. 61/388,782 filed on October 1, 2010, U.S. Provisional Application Serial No. 61/394,963, filed on October 20, 2010, U.S. Provisional Application Serial No. 61/405,005, filed on October 20, 2010 and U.S. Provisional Application Serial No. 61/405,079, filed on October 20, 2010; the contents of each of which are hereby incorporated by reference in their entirety.

Statement of Government Support
Aspects of the present invention were made with the support of funding under federal grant numbers K23DK083551, R01HL068927, R01HL68890 and R01DK66368 from the National Institutes of Health. The United States Government has certain rights to this invention.

Background
Cystic Fibrosis (CF) is a life-shortening recessive genetic disorder that is caused by mutations in the cystic fibrosis transmembrane conductance regulator (CFTR) gene. However, there is substantial variability in the clinical phenotypes of cystic fibrosis patients, even among individuals with the exact same loss-of-function CFTR mutation.

Individuals with cystic fibrosis often exhibit drastically different secondary disease manifestations and may have highly variable lung disease severity. For example, 16-25% of cystic fibrosis patients suffer from Meconium Ileus (MI) (a severe intestinal obstruction), while roughly 30% develop CF related diabetes (CFRD) in adulthood and 5-7% acquire cystic fibrosis related liver disease (CFLD). Furthermore, while almost all individuals with cystic fibrosis suffer from progressive bronchopulmonary disease, the severity and rate of decline in lung function highly variable among individuals carrying the same CFTR mutation.

Thus, there is great need for compositions and methods for the identification of individuals with cystic fibrosis who are at increased risk of severe lung disease, MI and/or CFLD.
Summary

In some embodiments, provided herein are methods of identifying a subject as having increased risk of severe lung disease. In some embodiments, the methods include the step of detecting in a biological sample from the subject an allele of a single nucleotide polymorphism. In certain embodiments the allele of the single nucleotide polymorphism is a) a C allele at single nucleotide polymorphism rs12793173, b) an A allele at single nucleotide polymorphism rs1403543, c) a C allele at single nucleotide polymorphism rs9268905, d) a G allele at single nucleotide polymorphism rs4760506, e) a T allele at single nucleotide polymorphism rs12883884, f) an A allele at single nucleotide polymorphism rs2188164, g) a C allele at single nucleotide polymorphism rs1645366 or h) any combination thereof. In some embodiments the methods include the step of detecting in a biological sample from the subject a variant of a gene. In some embodiments the gene is EHF, APIP, MC3R, CASS4, AURKA, CBLN4, C20orf106 and/or CSTF1.

In some embodiments, provided herein are methods of identifying subject as a carrier of an allele of a single nucleotide polymorphism associated with severe lung disease. In some embodiments, the methods include the step of detecting in a biological sample from the subject an allele of a single nucleotide polymorphism. In certain embodiments the allele of the single nucleotide polymorphism is a) a C allele at single nucleotide polymorphism rs12793173, b) an A allele at single nucleotide polymorphism rs1403543, c) a C allele at single nucleotide polymorphism rs9268905, d) a G allele at single nucleotide polymorphism rs4760506, e) a T allele at single nucleotide polymorphism rs2188164, f) an A allele at single nucleotide polymorphism rs12883884, g) a C allele at single nucleotide polymorphism rs1645366 or h) any combination thereof. In some embodiments the methods include the step of detecting in a biological sample from the subject a variant in a gene. In some embodiments the gene is EHF, APIP, MC3R, CASS4, AURKA, CBLN4, C20orf106 and/or CSTF1.

In some embodiments, provided herein are methods of identifying a subject as having increased risk of having cystic fibrosis liver disease (CFLD). In certain embodiments, the methods include the step of detecting in a biological sample from the subject an allele of a single nucleotide polymorphism. In certain embodiments the allele of the single nucleotide polymorphism is a) a C allele at single nucleotide polymorphism
rs914232, b) a T allele at single nucleotide polymorphism rs2330183, c) a T allele at single nucleotide polymorphism rs2838956, d) a G allele at single nucleotide polymorphism rs1051266, e) a T allele at single nucleotide polymorphism rs4819130, f) a G allele at single nucleotide polymorphism rs3788190, g) a T allele at single nucleotide polymorphism rs2236483, h) a C allele at single nucleotide polymorphism rs2838950, i) a G allele at single nucleotide polymorphism rs3753019 or k) any combination thereof. In some embodiments the methods include the step of detecting in a biological sample from the subject a variant in a gene. In some embodiments the gene is SLC19A1 and/or COL18A1.

In some embodiments, provided herein are methods of identifying a subject as a carrier of an allele of a single nucleotide polymorphism associated with CFLD. In certain embodiments, the methods include the step of detecting in a biological sample from the subject an allele of a single nucleotide polymorphism. In certain embodiments the allele of the single nucleotide polymorphism is a) a C allele at single nucleotide polymorphism rs914232, b) a T allele at single nucleotide polymorphism rs2330183, c) a T allele at single nucleotide polymorphism rs2838956, d) a G allele at single nucleotide polymorphism rs1051266, e) a T allele at single nucleotide polymorphism rs4819130, f) a G allele at single nucleotide polymorphism rs2236483, h) a C allele at single nucleotide polymorphism rs2838950, i) a G allele at single nucleotide polymorphism rs3753019 or k) any combination thereof. In some embodiments the methods include the step of detecting in a biological sample from the subject a variant in a gene. In some embodiments the gene is SLC19A1 and/or COL18A1.

In some embodiments, provided herein are methods of identifying a subject as having increased risk of meconium ileus (MI). In some embodiments the methods include the step of detecting in a biological sample from the subject an allele of a single nucleotide polymorphism. In some embodiments the allele of the single nucleotide polymorphism is a) a C allele at single nucleotide polymorphism rs7512462, b) a G allele at single nucleotide polymorphism rs7415921, c) a G allele at single nucleotide polymorphism rs4077468, d) a T allele at single nucleotide polymorphism rs4077469, e) a G allele at single nucleotide polymorphism rs2047830, f) an A allele at single nucleotide polymorphism rs7419153, g) a T allele at single nucleotide polymorphism rs10179921, h) a T allele at single nucleotide
polymorphism rs4684689, i) an A allele at single nucleotide polymorphism rs17563 161, j) a T allele at single nucleotide polymorphism rs3788766, k) a C allele at single nucleotide polymorphism rs5905283, l) a G allele at single nucleotide polymorphism rs12839137 or k) any combination thereof. In some embodiments the methods include the step of detecting in a biological sample from the subject a variant in a gene. In some embodiments the gene is SLC26A9, SLC6A14, SLC9A3, ABCG8 and/or ATP2B2.

In some embodiments, provided herein are methods of identifying a subject as a carrier of an allele of a single nucleotide polymorphism associated with MI. In some embodiments the methods include the step of detecting in a biological sample from the subject an allele of a single nucleotide polymorphism. In some embodiments the allele of the single nucleotide polymorphism is a) a C allele at single nucleotide polymorphism rs7512462, b) a G allele at single nucleotide polymorphism rs7415921, c) a G allele at single nucleotide polymorphism rs4077468, d) a T allele at single nucleotide polymorphism rs4077469, e) a G allele at single nucleotide polymorphism rs12047830, f) an A allele at single nucleotide polymorphism rs7419153, g) a T allele at single nucleotide polymorphism rs179921, h) a T allele at single nucleotide polymorphism rs4684689, i) an A allele at single nucleotide polymorphism rs7563 161, j) a T allele at single nucleotide polymorphism rs3788766, k) a C allele at single nucleotide polymorphism rs5905283, l) a G allele at single nucleotide polymorphism rs12839137 or k) any combination thereof. In some embodiments the methods include the step of detecting in a biological sample from the subject a variant in a gene. In some embodiments the gene is SLC26A9, SLC6A14, SLC9A3, ABCG8 and/or ATP2B2.

In certain embodiments of the methods described herein, the subject lacks a wild-type CFTR gene, has or is suspected of having cystic fibrosis, is or is suspected of being a carrier of a mutated CFTR gene, and/or has at least one family member that has or is suspected of having cystic fibrosis.

In some embodiments, the methods described herein also includes the step of determining whether the biological sample lacks a wild-type CFTR gene. In certain embodiments, the methods described herein include the step of obtaining the biological sample from the subject.
In some embodiments of the methods described herein, the step of detecting includes performing a hybridization assay, an amplification assay and/or a nucleic acid sequencing assay.

In some embodiments of the methods described herein, the sample a tissue sample, a blood sample, a semen sample and/or a germ cell sample. In certain embodiments, the subject is a human adult, a human child, a human fetus, a human embryo or a human fertilized cell.

In some embodiments, described herein are methods of determining whether a test compound is a candidate therapeutic agent for reducing lung disease severity. In certain embodiments, the methods include a) contacting a cell with the test compound; and b) detecting the expression by the cell of a gene product of EHF, APIP, MC3R, CASS4, AURKA, CBLN4, C20orf106 and/or CSTF1.

In some embodiments, described herein are methods of determining whether a test compound is a candidate therapeutic agent for treating CFLD. In certain embodiments, the methods include a) contacting a cell with the test compound; and b) detecting the expression by the cell of a gene product of SLC19A1 and/or COL18A1.

In some embodiments, described herein are methods of determining whether a test compound is a candidate therapeutic agent for treating MI. In certain embodiments, the methods include a) contacting a cell with the test compound; and b) detecting the expression by the cell of a gene product of SLC26A9, SLC6A14, SLC9A3, ABCG8 and/or ATP2B2.

In some embodiments of the methods described herein, the gene product is an mRNA and/or a protein. In certain embodiments the gene product is linked to a detectable moiety. In some embodiments the expression of the gene product is detected by detecting the detectable moiety. In certain embodiments, the agent is a small molecule, a polypeptide, an antibody or an inhibitory RNA molecule.

In some embodiments, described herein are methods of reducing lung disease severity in a subject. In certain embodiments, the methods include administering to the subject a therapeutic agent that modulates the expression or activity of a gene product encoded by EHF, APIP, MC3R, CASS4, AURKA, CBLN4, C20orf106 and/or CSTF1.

In some embodiments, described herein are methods of treating and/or preventing CFLD in a subject. In certain embodiments, the methods include administering to the
subject a therapeutic agent that modulates the expression or activity of a gene product encoded by SLC19A1 and/or COL18A1.

In some embodiments, described herein are methods of treating and/or preventing MI in a subject. In certain embodiments, the methods include administering to the subject a therapeutic agent that modulates the expression or activity of a gene product encoded by SLC26A9, SLC6A14, SLC9A3, ABCG8 and/or ATP2B2.

In certain embodiments of the methods described herein, the subject lacks a wild-type CFTR gene, has or is suspected of having cystic fibrosis, is or is suspected of being a carrier of a mutated CFTR gene, and/or has at least one family member that has or is suspected of having cystic fibrosis. In certain embodiments, the agent is a small molecule, a polypeptide, an antibody and/or an inhibitory RNA molecule. In some embodiments, the agent reduces the expression or activity of the gene product.

Brief Description of the Figures

Figure 1 is a table that provides of exemplary SNPs.

Figure 2 is a table that provides the characteristics of patients enrolled by three studies in the North American CF Gene Modifier Consortium.

Figure 3 shows histograms of the Consortium lung phenotype for the three cystic fibrosis studies show similar average phenotypes. The phenotype mean is above zero due to a lower bound placed by the survival correction, as well as cohort effects of improving lung function. (a) The two designs using unrelated individuals. All of the patients in the Genetic Modifier Study (GMS) are F508del/F508del at CFTR. These patients were oversampled at extremes of an initial entry phenotype, in order to improve power, and the original severe/mild designations are colored separately. In contrast, the Canadian Consortium for Genetic Studies (CGS) is population based, representing a range of pancreatic insufficient CFTR genotypes. (b) Patients enrolled in the family-based Twin and Sibling Study (TSS) show a similar distribution of the Consortium lung phenotype as the population-based CGS.

Figure 4 shows Power analyses for genome-wide significance of \( P<5 \times 10^{-8} \) for GMS and CGS F508del/F508del (\( n=1978 \)). Each plotted point is the result of 1000 simulations using the phenotype-conditional sampling scheme described in Online Methods. (a) Results as a function of (i) the population effect size \( \rho^2 \), defined as the change
in average Consortium lung phenotype for each additional copy of the minor allele, and (ii)
minor allele frequency. The power depends essentially on the proportion of variation
explained, as illustrated in (b). The power ranges from 25%-90% to detect variants
explaining 1%-2% of population phenotype variation. Power for the combined scan of
GMS and all CGS patients is even higher, due to the larger sample size for that scan.

Figure 5 shows genome-wide Manhattan plots for the cystic fibrosis Consortium
lung function phenotype, combining the association evidence from GMS and CGS samples
across 570,725 SNPs. The black dashed line represents the Bonferroni threshold for
genome-wide \( \alpha=0.05 \), while the gray dashed line is the suggestive association threshold,
expected once per genome scan. SNPs are plotted in Mb relative to their position on each
chromosome (alternating gray and black) (a) Results from GMS (\( \alpha=137 \), all of whom are
\( F508\text{del}/F508\text{del} \)) combined with all of the CGS patients (\( \alpha=1357 \)). Seven regions reach
suggestive significance, (b) Results from the combined evidence of GMS (\( \alpha=137 \)) and the
CGS \( F508\text{del}/F508\text{del} \) (n=841). A region on chromosome 11p13 reaches genome-wide
significance (p=3.34 X 10^{-8}).

Figure 6 is a table that provides significant and suggestive association results for
GMS and CGS with replication values for TSS.

Figure 7 shows an alternative analysis of association evidence for the Consortium
lung phenotype shows consistent evidence for the 11p13 \( EHF/APIP \) region. Results from
the conditional likelihood approach described in Online Methods applied to GMS+CGS
\( F508\text{del}/F508\text{del} \). The black and gray dashed lines correspond to genome-wide and
suggestive significance, respectively.

Figure 8 is a table that provides the covariate effects for significant and suggestive
SNPs.

Figure 9 shows (a) Joint analysis of association evidence from GMS and all
patients from CGS and TSS shows that the 11p13 \( EHF/APIP \) region reaches genome-wide
significance for this population set, and chromosome 6 region near \( HLA-DRA \) on 6p21
nearly achieves genome-wide significance, (b) Joint analysis of association evidence from
GMS and the \( F508\text{del}/F508\text{del} \) from CGS and TSS shows striking evidence in the
\( EHF/APIP \) region (P=8.28X10^{-10} at rs568529). The black and gray dashed lines correspond
to genome-wide and suggestive significance, respectively.
**Figure 10** shows a plot of the association evidence in GMS and CGS *F508del/F508del* in the chromosome 11p13 *EHF/APIP* region (NCBI build 36, LocusZoom viewer).

**Figure 11** shows Conditional analysis of the chromosome 11 association result for the Consortium lung phenotype, GMS+CGS *F508del/F508del*. After conditioning on the most significant SNP rs12793173 from the genome scan the minimum *p*-value in the interval is 3.73x10^-5 at SNP rs286873. The conditional result was significant for the interval (Bonferroni *p*=0.0029 for the 77 SNPs annotated to the *EHF/APIP* interval). After another step of conditioning on rs286873 alone, and performing another regional scan, the original SNP rs12793173 remained the most significant.

**Figure 12** shows an illustrative Manhattan plot for association of the Consortium lung phenotype with genoCNV copy number, at 2544 loci variable (variant frequency >1%) in a combined analysis of GMS and all CGS patients. No region attained Bonferroni genome-wide significance threshold for the 2544 CNV loci. The model shown here is trait = sex + PCs + cij + cij *sex + acj *sex (Model ii in Online Methods). P-values were computed by comparing the full model to the reduced model with no copy number information, and combined across the two studies using Fisher's combined *p*-value method.

**Figure 13** shows genome-wide linkage scan for the Consortium lung phenotype in the family-based TSS cystic fibrosis study, adjusted for sex. A QTL with a genome-wide significant LOD=5.04 was found on 20q13.2. SNP used in the linkage panel are plotted in cM relative to their position on each chromosome (alternating gray and black).

**Figure 14** shows correlation plots of the lung function measures of sibling pairs who do not share either parental allele of rs4811626 on chromosome 20 (Identity by descent or IBD 0; 117 pairs), share one allele (IBD1; 248 pairs) or share both parental alleles (IBD2; 116 pairs). The IBD status could not be assigned for 5 sibling pairs. Correlation was calculated when sex and BMI were used as covariates. Correlation by IBD status was as follows: IBD0: r=0.1762, r^2=0.0310; IBD1: r= 0.4136 r^2=0.171 and IBD2: r= 0.6542, r^2=0.428. The contribution of the linked region to variation was estimated by subtracting the correlation in siblings who are IBD 0 from the correlation in sibling who are IBD 2 (0.6542 - 0.1762 = 0.478). Similar estimates were obtained using variance components methods in Merlin (0.499) and SOLAR (0.496).
**Figure 15** shows regional analysis of the QTL on chromosome 20q13 (a) A detailed chromosome 20 linkage plot for the Consortium lung phenotype in the TSS study, with covariates sex (essentially the same result as for no covariates) and with covariates sex and BMI. (b) Association evidence from the GMS and CGS F508del/F508del patients, in the 1- LOD support interval provided by TSS.

**Figure 16** is a table that provides combined association and linkage-weighted FDR q-values and genome-wide ranks for SNPs with WFDR 1-values of genome wide significance (<0.05).

**Figure 17** shows a weighted false-discovery rate analysis (WFDR) is used to provide combined linkage evidence from TSS with association evidence from GMS+CGS F508del/F508del. A Manhattan plot portrays the ranks of the SNP WFDR g-values, with the rank for these data corresponding to $q=0.05$ shown as a dashed line. SNPs with g-values less than 0.05 are plotted above the dotted line. Note that the 1p13 and 20q13 regions are both significant, with WFDR g-values of 0.028 and 0.015, respectively.

**Figure 18** shows a proposed mechanism of stellate cell activation and fibrogenesis in CFLD. Hepatic stellate cell activation is amplified by the combined stimulus of hepatocytes, cholangiocytes and multiple stimuli that may reflect genetic factors (only a few of the known mediators are shown).

**Figure 19** shows a Manhattan plot of SNP p-values in CFLD. The top dotted line is threshold for genome-wide significance and the bottom dotted line is "suggestive" (i.e., only expect one SNP above this line by chance).

**Figure 20** shows SNPs on chromosome 21 (Chr 21) that associate with CFLD, and plotted relative to known genes and recombination rates.

**Figure 21** is a table that provides details of the CF consortium participants of the MI study.

**Figure 22** shows a QQ-plot of the MI genome-wide association (GWAS) analysis performed via a GEE model.

**Figure 23** is a table that provides the sex-specific results for rs3788766.

**Figure 24** shows a regional plot of the association evidence for MI around the solute carrier protein gene, SLC6A14, on chromosome X.
Figure 25 is a set of tables from the MI study that provide (a) GWAS results for all CF patients, (b) GWAS results for AF508 Homozygous CF patients, and (c) OR estimates for all SNPs in (a) with a q-value <0.05.

Figure 26 is a table that provides a list of genes that encode proteins that localize to the apical plasma membrane.

Figure 27 shows a QQ-plot of SNPs in apical plasma membrane genes (A) and SNPs in other genes (B).

Figure 28 shows a QQ-plot of SNPs from apical plasma membrane genes (A) and nuclear envelope genes (B) based on observed data and permuted phenotype data under the null hypothesis of no association.

Figure 29 shows a trace plot of SNP association in the CFTR region (A) and a regional plot of the association evidence in and around CFTR (B).

Figure 30 is a table that provides the results of the GWAS-HD analysis for MI.

Figure 31 shows a regional plot of the MI association evidence in and around SLC26A9.

Figure 32 shows the polymorphism in the CEBPB binding site of the SLC6A14 promoter region.

Figure 33 shows a QQ-plot from the GWAS for the combined statistic to detect pleiotropic effects for MI and lung disease.

Figure 34 is a table that provides the top 16 SNPs (p-value ≤ 10^-5) according to the genome-wide MI association qq-plot shown in Figure 33.

Figure 35 shows that MI GWAS leads to genome-wide significant SNPs. (a) Genome-wide Manhattan plot for MI. The black solid line represents the genome-wide significance threshold (P value<5x10^-8), and the black dashed line is the suggestive association threshold, expected once per genome scan. (b) Regional plot for SLC26A9. LocusZoom viewer was used to generate and display the association evidence around SLC26A9 based on NCBI Build 36/hg18. (c) Regional plot for SLC6A14.

Figure 36 is a table that provides information regarding Mi-associated SNPs in SLC26A9 and SLC6A14.

Figure 37 is a table that provides the sex-specific results for rs3788766 and rs5905283 in SLC6A14.
Figure 38 shows genome-wide MI association results with and without adjusting for the effect of CFTR. The x-axis shows the association P values (on the -log 10 scale) of the original GWAS with the site covariate but without adjusting for the effect of CFTR as in Figure 1a; the y-axis shows the association P values with both the site covariate and the CFTR covariate for which Phe508del/Phe508del genotype is coded as 1 and Phe508del/Other or Other/Other genotypes are coded as 0. SNPs within 155 kb of CFTR have been removed from this figure, and the SNPs at the bottom-left that have some noticeable discrepancy between the two sets of analyses are the SNPs that are in LD with CFTR.

Figure 39 is a table that provides Mi-association results for SNPs in SLC6A14 and SLC26A9 with and without adjustment for CFTR.

Figure 40 shows that the apical membrane hypothesis identifies genes associated with MI. (a) QQ-plot of the apical SNPs in the discovery sample. (b) Statistical significance of the apical membrane hypothesis in the discovery sample. Statistical significance (permutation P = 0.0002) was established via a sum statistic, summing the association evidence (Wald $\chi^2$ statistic) over all the 3,814 SNPs with observed sum statistic. (c) QQ-plot of the apical SNPs in the replication sample. (d) Statistical Significance of the apical membrane hypothesis in the replication sample (permutation P = 22/1,000 = 0.022).

Figure 41 is a table that provides gene-based and Lasso association results for SLC6A14 and 157 Apical Genes.

Figure 42 shows a GWAS-HD flow chart.

Figure 43 is a table that provides ranked SNPs with MI association with q values <0.05 from GWAS or GWAS-HD.

Figure 44 shows assessment of the nuclear envelope null hypothesis, (a) QQ-plot of the nuclear envelope gene SNPs in the discovery sample. (b) Statistical significance of the nuclear envelope hypothesis in the discovery sample.

Detailed Description

I. General

Despite the fact that cystic fibrosis (CF) is considered a "monogenic" recessive disease caused by the mutation of the CFTR gene, there is substantial variability in CF
clinical phenotype, even among individuals carrying the exact same CFTR mutations. Provided herein are genetic markers (e.g., SNP alleles and gene variants) associated with increased risk of severe lung disease, cystic fibrosis related liver disease (CFLD) and/or meconium ileus (MI) in individuals with cystic fibrosis. As described herein, such SNPs and gene variants are useful, for example, in methods of identifying a subject (e.g., a subject who has or is suspected of having CF) as having an increased risk of severe lung disease, CFLD and/or MI. Such genetic markers are also useful for the identification of individuals who carry genetic modifiers of cystic fibrosis clinical phenotype, the identification of novel therapeutic agents and for the treatment of lung disease, CFLD and/or MI.

Also described herein are therapeutic targets which can be modulated in order to treat and/or prevent cystic fibrosis, severe lung disease, CFLD and/or MI. Such therapeutic targets are also useful for the identification of novel therapeutic agents for the treatment of cystic fibrosis, severe lung disease, CFLD and/or MI.

II. Definitions

In order that the present invention may be more readily understood, certain terms are first defined. Additional definitions are set forth throughout the detailed description.

The articles "a" and "an" are used herein to refer to one or to more than one (i.e., to at least one) of the grammatical object of the article. By way of example, "an element" means one element or more than one element.

As used herein, the term "administering" means providing a pharmaceutical agent or composition to a subject, and includes, but is not limited to, administering by a medical professional and self-administering.

The term "agent" is used herein to denote a chemical compound, a small molecule, a mixture of chemical compounds, a biological macromolecule (such as a nucleic acid, an antibody, a protein or portion thereof), or an extract made from biological materials such as bacteria, plants, fungi, or animal (particularly mammalian) cells or tissues. Agents may be identified as having a particular activity (e.g., modulating a therapeutic target) by screening assays described herein below. The activity of such agents may render them suitable as a "therapeutic agent" which is a biologically, physiologically, or pharmacologically active substance (or substances) that acts locally or systemically in a subject.
As used herein, an "allele" refers to one of two or more alternative forms of a nucleotide sequence at a given position (locus) on a chromosome. An individual can be heterozygous or homozygous for any allele of described herein.

The term "altered level of expression" or "modulated expression" of a gene product (e.g., a therapeutic target described herein) refers to an expression level of a gene product in a cell or sample that has been contacted with an agent that is greater or less than the expression level of the same gene product a control cell or sample (e.g., a cell or sample of the same type that has not been contacted with the agent or that has been contacted with a placebo agent).

The term "altered activity" or "modulated activity" of a gene product (e.g., a therapeutic target described herein) refers to an activity level of a gene product in a cell or sample that has been contacted with an agent that is greater or less than the activity level of the same gene product a control cell or sample (e.g., a cell or sample of the same type that has not been contacted with the agent or that has been contacted with a placebo agent).

Altered activity may be the result of, for example, altered mRNA level, altered protein level, altered structure, altered ligand binding, and interference with protein-protein interactions.

As used herein, the term "antibody" includes full-length antibodies and any antigen binding fragment (i.e., "antigen-binding portion") or single chain thereof. The term "antibody" includes, but is not limited to, a glycoprotein comprising at least two heavy (H) chains and two light (L) chains inter-connected by disulfide bonds, or an antigen binding portion thereof. Antibodies may be polyclonal or monoclonal; xenogeneic, allogeneic, or syngeneic; or modified forms thereof (e.g., humanized, chimeric).

As used herein, the term "cystic fibrosis" or "CF" describes a recessive genetic disorder that manifests in individuals who have two bona fide mutations in trans in the cystic fibrosis transmembrane conductance regulator (CFTR) gene. The mRNA and protein sequences of wild-type CFTR are provided at GenBank® accession numbers NM_000492.3 and NP_000483.3, respectively. Cystic fibrosis causing mutations in the CFTR gene are well known in the art. The most common CFTR mutation is the AF508 mutation.

As used herein, the phrases "gene product" and "product of a gene" refers to a substance encoded by a gene and able to be produced, either directly or indirectly, through
the transcription of the gene. The phrases "gene product" and "product of a gene" include RNA gene products (e.g. mRNA), DNA gene products (e.g. cDNA) and polypeptide gene products (e.g. proteins).

The terms "increased risk" or "increased likelihood" as well as "decreased risk" or "decreased likelihood" as used herein define the level of risk or the likelihood that a subject has or will develop severe lung disease, CFLD, or MI, as compared to a control subject that does not carry one or more of the alleles of a single nucleotide polymorphism or the mutated genes described herein.

As used herein, a "marker", "genetic marker," "polymorphic marker" or "polymorphism" is a genomic DNA sequence associated with and individual at increased risk for severe lung disease, CFLD or MI. Each polymorphic marker has at least two sequence variations characteristic of particular alleles at the polymorphic site. Thus, genetic association to a polymorphic marker implies that there is association to at least one specific allele of that particular polymorphic marker. The marker can comprise any allele of any variant type found in the genome, including SNPs, mini- or microsatellites, translocations and copy number variations (insertions, deletions, duplications). Polymorphic markers can be of any measurable frequency in the population.

As used herein, the term "modulation" refers to up regulation (i.e., activation or stimulation), down regulation (i.e., inhibition or suppression) of the expression of a gene product, of a biological activity, or the two in combination or apart.

The term "pharmaceutically acceptable carrier" is art-recognized and refers to a pharmaceutically-acceptable material, composition or vehicle, such as a liquid or solid filler, diluent, excipient, solvent or encapsulating material, involved in carrying or transporting any subject composition or component thereof from one organ, or portion of the body, to another organ, or portion of the body. Each carrier must be "acceptable" in the sense of being compatible with the subject composition and its components and not injurious to the patient. Some examples of materials which may serve as pharmaceutically acceptable carriers include: (1) sugars, such as lactose, glucose and sucrose; (2) starches, such as corn starch and potato starch; (3) cellulose, and its derivatives, such as sodium carboxymethyl cellulose, ethyl cellulose and cellulose acetate; (4) powdered tragacanth; (5) malt; (6) gelatin; (7) talc; (8) excipients, such as cocoa butter and suppository waxes; (9) oils, such as peanut oil, cottonseed oil, safflower oil, sesame oil, olive oil, corn oil and
soybean oil; (10) glycols, such as propylene glycol; (11) polyols, such as glycerin, sorbitol, mannitol and polyethylene glycol; (12) esters, such as ethyl oleate and ethyl laurate; (13) agar; (14) buffering agents, such as magnesium hydroxide and aluminum hydroxide; (15) alginic acid; (16) pyrogen-free water; (17) isotonic saline; (18) Ringer's solution; (19) ethyl alcohol; (20) phosphate buffer solutions; and (21) other non-toxic compatible substances employed in pharmaceutical formulations.

"Sample," "tissue sample," "subject sample," or "biological sample" each refers to a collection of cells obtained from a tissue of a subject. The source of the tissue sample may be solid tissue, as from a fresh, frozen and/or preserved organ, tissue sample, biopsy, or aspirate; blood or any blood constituents, serum, blood; bodily fluids such as cerebral spinal fluid, amniotic fluid, peritoneal fluid or interstitial fluid, urine, saliva, stool, tears; or cells from any time in gestation or development of the subject. The tissue sample may contain compounds that are not naturally intermixed with the tissue in nature such as preservatives, anticoagulants, buffers, fixatives, nutrients, antibiotics or the like.

A "Single Nucleotide Polymorphism" or "SNP" is a DNA sequence variation occurring when a single nucleotide at a specific location in the genome differs between members of a species or between paired chromosomes in an individual. Most SNP polymorphisms have two alleles. Each individual is in this instance either homozygous for one allele of the polymorphism (i.e. both chromosomal copies of the individual have the same nucleotide at the SNP location), or the individual is heterozygous (i.e. the two sister chromosomes of the individual contain different nucleotides). The SNP nomenclature as reported herein refers to the official Reference SNP (rs) ID identification tag as assigned to each unique SNP by the National Center for Biotechnological Information (NCBI). A SNP allele can be describe based on the sequence of its forward strand or the sequence of its reverse strand. For example, a SNP that has either A or G alleles on its forward strand will have either T or C alleles, respectively, on its reverse strand. The SNP alleles are described herein according to their forward strand sequence. Exemplary SNPs are provided in Figure 1, along with their forward strand flanking sequences and their chromosomal position.

The term "small molecule" is art-recognized and refers to a composition which has a molecular weight of less than about 2000 amu, or less than about 1000 amu, and even less than about 500 amu. Small molecules may be, for example, nucleic acids, peptides, polypeptides, peptide nucleic acids, peptidomimetics, carbohydrates, lipids or other organic
(carbon containing) or inorganic molecules. Many pharmaceutical companies have extensive libraries of chemical and/or biological mixtures, often fungal, bacterial, or algal extracts, which can be screened with any of the assays described herein. The term "small organic molecule" refers to a small molecule that is often identified as being an organic or medicinal compound, and does not include molecules that are exclusively nucleic acids, peptides or polypeptides.

As used herein, the terms "subject" and "subjects" refer to an animal, e.g., a mammal including a non-primate (e.g., a cow, pig, horse, donkey, goat, camel, cat, dog, guinea pig, rat, mouse, sheep) and a primate (e.g., a monkey, such as a cynomolgous monkey, gorilla, chimpanzee and a human). In some embodiments, the subject may be a human adult, a human child, a human fetus, a human embryo and/or a human fertilized cell.

As used herein, the term "target" or "therapeutic target" are used interchangeably and refer to a gene product whose activity and/or expression can be modulated in order to treat and/or prevent a disease or disorder.

The phrases "therapeutically-effective amount" and "effective amount" as used herein means that amount of a therapeutic agent which is effective for producing some desired therapeutic effect in at least a sub-population of cells in an animal at a reasonable benefit/risk ratio applicable to any medical treatment.

"Treating" a disease in a subject or "treating" a subject having a disease refers to subjecting or exposing the subject to a pharmaceutical treatment, e.g., the administration of a drug, such that at least one symptom of the disease is decreased or prevented from worsening.

As used herein, the terms "variant of a gene," "gene variant," "mutation of a gene" and "gene mutation" are used interchangeably and refer to a particular allele of a gene described herein that is associated with increased risk for a disease or disorder. The variant may be functional or non-functional. The variant or mutation may be the gene allele that is less prevalent among the general population, but, in some instances, the variant or mutation may be the allele that is more prevalent among the general population.
III. SNPs and Gene Variants Associated with Lung Disease Severity

Lung disease is the major source of morbidity and mortality for patients afflicted with cystic fibrosis (CF), a recessive disorder caused by mutations in the cystic fibrosis transmembrane conductance regulator (CFTR) gene. The identification of CFTR and its disease-causing mutations provided substantial insight into the molecular pathophysiology of CF, but allelic variation in CFTR does not explain the wide variation in severity of lung disease. Therefore, identification of the genetic modifiers would increase our understanding of CF disease progression, suggest new targets for intervention and could identify potential mechanisms for variation in lung function in CF, as well as, for common chronic respiratory diseases such as COPD.

Provided herein are predictive alleles of single nucleotide polymorphisms associated with an increased risk of severe lung disease. As described herein, such alleles are therefore genetic markers of an increased risk of severe lung disease and are useful, for example, in the methods described herein for the identification of individuals as having increased risk of severe lung disease. These alleles include, but are not limited to, a C allele at single nucleotide polymorphism rs12793173, an A allele at single nucleotide polymorphism rs1403543, a C allele at single nucleotide polymorphism rs9268905, a G allele at single nucleotide polymorphism rs4760506, a T allele at single nucleotide polymorphism rs12883884, an A allele at single nucleotide polymorphism rs1645366. In certain embodiments, combinations comprising one or more of the predictive alleles are used in the methods described herein. These single nucleotide polymorphisms are identified by a reference number that can be found in the publicly available GenBank® database, well known to those of skill in the art.

Also provided herein are genes associated with severe lung disease in individuals with cystic fibrosis. Like the alleles of single nucleotide polymorphisms described above, variants of such genes are genetic markers of an increased risk of severe lung disease and are therefore useful, for example, in the methods described herein for the identification of individuals as having increased risk of severe lung disease. Furthermore, such the products of such genes are also therapeutic targets for the treatment of severe lung disease. As such, they are useful, for example, in the methods described herein for the treatment of severe lung disease and in the methods described herein for the identification of therapeutic agents.
for the treatment of severe lung disease. In some embodiments, agents that modulate the activity and/or expression of such therapeutic targets are identified as candidate therapeutic agents useful in the reduction in the severity of lung disease in individuals with cystic fibrosis. Furthermore, in certain embodiments, modulation of the activity and/or expression of such therapeutic targets are used to lung disease severity in individuals with cystic fibrosis.

Severe lung disease associated genes provided herein include EHF, APIP, MC3R, CASS4, AURKA, CBLN4, C20orf106 and CSTF1. The following GenBank® Database Accession numbers provide the wild-type sequences for the mRNA and protein encoded by each of these genes:

EHF: NP_001193544.1 (protein); NM_001206615.1 (mRNA).
APIP: NP_057041.2 (protein); NM_01206615.1 (mRNA).
AURKA: NP_940835.1 (protein); NM_01206615.1 (mRNA).
CBLN4: NP_542184.1 (protein); NM_080617.4 (mRNA).
C20orf106: NP_001012989.2 (protein); NM_001012971.3 (mRNA).
CSTF1: NP_001028693.1 (protein); NM_001033521.1 (mRNA).

In certain embodiments, the gene variants described herein include any mutation which modulates the activity and/or expression of a product of the severe lung disease associated gene. For example, such mutations can be insertion, deletion and/or substitution mutations. In certain embodiments, the mutation is a loss of function mutation. In some embodiments the mutation is a frame shift mutation and/or a truncation. In certain embodiments the variant is not a mutation to a coding portion of the severe lung disease associated gene, but rather to a transcription control element operably linked to the gene. For example, in some embodiments, the mutation is to a promoter or enhancer of the severe lung disease associated gene.

A subject can be homozygous or heterozygous for one or more of the genetic markers described herein, in any combination. For example, a subject can be homozygous for one marker and heterozygous for another marker and such homozygous and/or heterozygous markers can be present in any combination in a subject.
IV. SNPs and Gene Variants Associated with Cystic Fibrosis Liver Disease (CFLD)

There is currently no way to identify children at risk for CFLD and the molecular pathogenesis is not currently understood. Because CFLD develops early in life (median age -10 years), "preventive" intervention in this irreversible process would need to be undertaken early. Furthermore, since only 5% of CF patients develop CFLD, there is currently no feasible study design to test potential therapies.

Provided herein are predictive alleles of single nucleotide polymorphisms associated with an increased risk of CFLD. As described herein, such alleles are therefore genetic markers of an increased risk of CFLD and are useful, for example, in the methods described herein for the identification of individuals as having increased risk of CFLD. These alleles include, but are not limited to, a C allele at single nucleotide polymorphism rs914232, a T allele at single nucleotide polymorphism rs2330183, an A allele at single nucleotide polymorphism rs2838956, a G allele at single nucleotide polymorphism rs1051266, a T allele at single nucleotide polymorphism rs4819130, a G allele at single nucleotide polymorphism rs3788190, a T allele at single nucleotide polymorphism rs2236483, a C allele at single nucleotide polymorphism rs2838950, a G allele at single nucleotide polymorphism rs12483377, and/or a C allele at single nucleotide polymorphism rs3753019. In certain embodiments, combinations of the predictive alleles are used in the methods described herein. These single nucleotide polymorphisms are identified by a reference number that can be found in the publicly available GenBank® database, well known to those of skill in the art.

Also provided herein are genes associated with CFLD. Like the alleles of single nucleotide polymorphisms described above, variants to such genes are genetic markers of an increased risk of CFLD and are therefore useful, for example, in the methods described herein for the identification of individuals as having increased risk of CFLD. Furthermore, the products of such genes are also therapeutic targets for the treatment of CFLD. As such, they are useful, for example, in the methods described herein for the treatment of CFLD and in the methods described herein for the identification of therapeutic agents for the treatment of CFLD. In some embodiments, agents that modulate the activity and/or expression of such therapeutic targets are identified as candidate therapeutic agents for the prevention and/or treatment of CFLD. Furthermore, in certain embodiments, modulation of
the activity and/or expression of the therapeutic targets described herein are used to prevent and/or treat CFLD.

CFLD associated genes provided herein include SLC19A1 and COL18A1. The following GenBank® Database Accession numbers provide the wild-type sequences for the mRNA and protein encoded by each of these genes:

<table>
<thead>
<tr>
<th>Gene</th>
<th>Accession Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLC19A1</td>
<td>NP_919231.1 (protein); NM_194255.1 (mRNA).</td>
</tr>
<tr>
<td>COL18A1</td>
<td>Isoform 1: NP_085059.2 (protein); NM_030582.3 (mRNA).</td>
</tr>
<tr>
<td></td>
<td>Isoform 2: NP_56971 1.2 (protein); NMJ30444.2 (mRNA).</td>
</tr>
<tr>
<td></td>
<td>Isoform 3: NP 569712.2 (protein); NMJ30445.2 (mRNA).</td>
</tr>
</tbody>
</table>

In certain embodiments, the gene variants described herein include any mutation which modulates the activity and/or expression of a product of the CFLD associated gene. For example, such mutations can be insertion, deletion and/or substitution mutations. In certain embodiments, the mutation is a loss of function mutation. In some embodiments the mutation is a frame shift mutation and/or a truncation. In certain embodiments the variant is not a mutation to a coding portion of the CFLD associated gene, but rather to an transcription control element operably linked to the gene. For example, in some embodiments the mutation is to a promoter or enhancer of the CFLD associated gene.

A subject can be homozygous or heterozygous for one or more of the genetic markers described herein, in any combination. For example, a subject can be homozygous for one marker and heterozygous for another marker and such homozygous and/or heterozygous markers can be present in any combination in a subject.

V. SNPs and Gene Variants Associated with Meconium Ileus

Meconium ileus (MI), a type of intestinal blockage, is seen in 16-25% of CF patients at birth with an equal sex ratio, but is otherwise rare such that presence of MI at birth is highly indicative of CF. Furthermore, MI almost exclusively occurs in patients with two severe CF-causing CFTR mutations (~87% of all patients). Although food digestion does not occur in utero, various facets of the gastrointestinal tract do begin to function including production of digestive enzymes, with clearance of material being essential immediately after birth. CFTR is expressed in various segments of the small and large intestine. With the loss of CFTR, the meconium (or first stool in the newborn) is altered as the intestinal mucus secretions that begin in utero are abnormally sticky and adherent...
leading to a blockage of the latter portion of the small intestine. The proximal ileum can be
enlarged and the subsequent distal ileum and the colon may appear collapsed. The
obstructions are dense material comprised of a mixture of bile salts, bile acids and debris
that is typically shed from the intestinal mucosa during the fetal period. Intestinal
obstruction due to MI will be evident as early as 24-48 hours after birth with distention,
vomiting and failure to pass meconium. Intervention to remove the blockage is required
immediately, via an enema procedure or by surgical intervention.

Provided herein are predictive alleles of single nucleotide polymorphisms
associated with an increased risk of MI. As described herein, such alleles are therefore
genetic markers of an increased risk of MI and are useful, for example, in the methods
described herein for the identification of individuals as having increased risk of MI. These
alleles include, but are not limited to, a C allele at single nucleotide polymorphism
rs7512462, a G allele at single nucleotide polymorphism rs7415921, a G allele at single
nucleotide polymorphism rs4077468, a T allele at single nucleotide polymorphism
rs4077469, a G allele at single nucleotide polymorphism rs12047830, an A allele at single
nucleotide polymorphism rs7419153, a T allele at single nucleotide polymorphism
rs179921, a T allele at single nucleotide polymorphism rs4684689, an A allele at single
nucleotide polymorphism rs17563161, a T allele at single nucleotide polymorphism
rs3788766, a C allele at single nucleotide polymorphism rs5905283 and/or a G allele at
single nucleotide polymorphism rs12839137. In certain embodiments, combinations of the
predictive alleles are used in the methods described herein. These single nucleotide
polymorphisms are identified by a reference number that can be found in the publicly
available GenBank® database, well known to those of skill in the art.

Also provided herein are genes associated with MI. Like the alleles of single
nucleotide polymorphisms described above, variants of such genes are genetic markers of
an increased risk of MI and are therefore useful, for example, in the methods described
herein for the identification of individuals as having increased risk of MI. Furthermore, the
products of such genes are therapeutic targets for the treatment of MI. As such, they are
useful, for example, in the methods described herein for the treatment of MI and in the
methods described herein for the identification of therapeutic agents for the treatment of
MI. In some embodiments, agents that modulate the activity and/or expression of such
therapeutic targets are identified as candidate therapeutic agents for the prevention and/or
treatment of MI. Furthermore, in certain embodiments, modulation of the activity and/or expression of such therapeutic targets are used to prevent and/or treat MI.

MI associated genes provided herein include SLC26A9, SLC6A14, SLC9A3, ABCG8 and ATP2B2. The following GenBank® Database Accession numbers provide the wild-type sequences for the mRNA and protein encoded by each of these genes:

SLC26A9:  
NP_443166.1 (protein); 1.NM_052934.3 (mRNA).

SLC6A14:  
NP_009162.1 (protein); NM_007231.3 (mRNA).

SLC9A3:  
NP_004165.2 (protein); NM_004174.2 (mRNA).

ABCG8:  
NP_071882.1 (protein); NM_022437.2 (mRNA).

ATP2B2:  
NP_00100133.1 (protein); NM_00100133.2 (mRNA).

In certain embodiment, the gene variants described herein include any mutation which modulates the activity and/or expression of a product of the MI associated gene. For example, such mutations can be insertion, deletion and/or substitution mutations. In certain embodiments, the mutation is a loss of function mutation. In some embodiments the mutation is a frame shift mutation and/or a truncation. In certain embodiments the variant is not a mutation to a coding portion of the MI associated gene, but rather to a transcription control element operably linked to the gene. For example, in some embodiments the mutation is to a promoter or enhancer of the MI associated gene.

A subject can be homozygous or heterozygous for one or more of the genetic markers described herein, in any combination. For example, a subject can be homozygous for one marker and heterozygous for another marker and such homozygous and/or heterozygous markers can be present in any combination in a subject.

VI. Identification of Subjects with Increased Risk of Severe Lung Disease, CFLD and MI

Described herein are methods of identifying a subject who has an increased risk of severe lung disease, CFLD and/or MI. In certain embodiments, the method includes the step of detecting in a biological sample from a subject a genetic marker described herein. In some embodiments, the genetic marker is an allele of a single nucleotide polymorphism that is associated with severe lung disease, CFLD and/or MI. In some embodiments, the genetic marker is a variant of a gene that is associated with severe lung disease, CFLD and/or MI. In some embodiments, the method comprises a combination of any one or more genetic markers described herein are detected. In general, if the genetic marker is detected
in the biological sample, the subject from whom the biological sample was obtained has an increased risk of severe lung disease, CFLD and/or MI. In certain embodiments, the subject has or is suspected of having cystic fibrosis. For example, in certain embodiments, the subject lacks a wild-type CFTR gene. In some embodiments, the subject has at least one family member that has or is suspected of having cystic fibrosis. In some embodiments, the method also includes the step of obtaining the biological sample from the subject.

In certain embodiments, the methods described herein also include the step of detecting mutated and/or wild-type CFTR in the sample. As described herein, individuals with cystic fibrosis carry mutations in both copies of their CFTR gene. Thus, in certain embodiments, the methods described herein determine both whether the subject has cystic fibrosis and whether the subject is at increased risk of severe lung disease, CFLD and/or MI.

Mutation of a single CFTR gene in a subject results in the subject being a carrier of the cystic fibrosis mutation. When two CFTR mutation carriers have a child, there is a one in four chance that the child will have cystic fibrosis. It is therefore desirable to know both whether an individual is a carrier of a cystic fibrosis causing mutation, but also whether an individual is a carrier of a genetic marker described herein.

Thus, described herein are methods of identifying a subject as a carrier of a genetic marker associated with severe lung disease, CFLD and/or MI. In certain embodiments, the method includes the step of detecting in a biological sample from a subject a genetic marker described herein. In some embodiments, the genetic marker is an allele of a single nucleotide polymorphism that is associated with severe lung disease, CFLD and/or MI. In some embodiments, the genetic marker is a variant of a gene that is associated with severe lung disease, CFLD and/or MI. In some embodiments, a combination of the genetic markers described herein are detected. In general, if the genetic marker is detected in the biological sample, the subject from whom the biological sample was obtained is a carrier of a genetic marker associated with severe lung disease, CFLD and/or MI. In some embodiments, the subject has at least one family member that has or is suspected of having cystic fibrosis. In certain embodiments, the subject is a carrier of a CFTR mutation. In some embodiments, the method also includes the step of obtaining the biological sample from the subject.
In some embodiments, of the methods described herein, the subject will be a human child or a human adult. In some embodiments, the subject will be an infant. However, in certain embodiments, the subject is not limited to being a fully developed human. Thus, in some embodiments, the subject will be a human fetus, a human embryo and/or a human fertilized cell.

Any type of biological sample that contains genetic material can be used in the methods described herein. Thus, for example, in some embodiments the sample is a cell, a body fluid, a swabbing, a tissue sample, a blood sample and/or a germ cell sample.

Any method known in the art can be used to detect the genetic markers described herein and/or the CFTR gene. Thus, in certain embodiments, the detecting step includes performing a hybridization assay (e.g., SNP or gene microarrays, dynamic allele-specific hybridization (DASH), TaqMAN, HPA, scorpion probes and molecular beacons), performing a nucleic acid amplification assay (e.g., PCR, LCR, TMA, SDA, NASBA, BDA, 3SR, RCR, etc.) and/or performing a nucleic acid sequencing assay.

In some embodiments, analysis of the nucleic acid can be carried by amplification of the region of interest according to amplification protocols well known in the art (e.g., polymerase chain reaction, ligase chain reaction, strand displacement amplification, transcription-based amplification, self-sustained sequence replication (3SR), QP replicase protocols, nucleic acid sequence-based amplification (NASBA), repair chain reaction (RCR) and boomerang DNA amplification (BDA), etc.). The amplification product can then be visualized directly in a gel by staining or the product can be detected by hybridization with a detectable probe. When amplification conditions allow for amplification of all allelic types of a genetic marker, the types can be distinguished by a variety of well known methods, such as hybridization with an allele-specific probe, secondary amplification with allele-specific primers, by restriction endonuclease digestion, and/or by electrophoresis. Thus, also provided herein are oligonucleotides for use as primers and/or probes for detecting and/or identifying genetic markers according to the methods described herein.

Additional methods for detecting the genetic markers described herein include sequencing, high performance liquid chromatography (HPLC), restriction enzyme analysis (e.g., restriction fragment length polymorphism or RFLP), hybridization, matrix assisted laser desorption/ionization-time of flight mass spectrometry (MALDI-TOF-MS), etc., all of
which are well known protocols for analyzing a nucleotide sequence and detecting genetic markers. The methods described herein can be carried out by using any assay or procedure that can interrogate a nucleic acid sequence.

In some embodiments, detecting can be carried out by an amplification reaction and single base extension, and in further embodiments, the product of the amplification reaction and single base extension can be spotted on a silicone chip according to methods well known in the art.

Prior to analyzing the sample, it may be necessary to process the sample to yield a form acceptable for analysis. For example, the nucleic acid (e.g. genomic DNA) may be extracted from the sample using techniques well-established in the art including chemical extraction techniques utilizing phenol-chloroform, guanidine-containing solutions, or CTAB-containing buffers. As well, as a matter of convenience, commercial DNA extraction kits are also widely available from laboratory reagent supply companies, including for example, the QIAamp DNA Blood Minikit available from QIAGEN (Chatsworth, CA), or the Extract-N-Amp blood kit available from Sigma (St. Louis, MO).

In certain embodiments, also provided herein is a kit comprising reagents to detect one or more of the markers described herein in a biological sample from a subject. Such a kit can comprise primers, probes, primer/probe sets, reagents, buffers, etc., as would be known in the art, for the detection of the genetic markers described herein in a biological sample from a subject. For example, a primer or probe can comprise a contiguous nucleotide sequence that is complementary (e.g., fully (100%) complementary or partially (50%, 60%, 70%, 80%, 90%, 95%, etc.) complementary) to a region comprising a marker described herein. In particular embodiments, a kit described herein can comprise primers and probes that allow for the specific detection of at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19 or 20 of the markers described herein. Such a kit can further comprise blocking probes, labeling reagents, blocking agents, restriction enzymes, antibodies, sampling devices, positive and negative controls, etc., as would be well known to those of skill in the art.

Also provided herein are methods of identifying an effective and/or appropriate (i.e., for a given subject's particular condition or status) treatment regimen for a subject with increased risk of severe lung disease, CFLD and/or MI, that includes detecting one or more of the genetic markers described herein in the subject, wherein the one or more
genetic markers are further statistically correlated with an effective and/or appropriate treatment regimen for cystic fibrosis, severe lung disease, CFLD and/or MI according to protocols as described herein and as are known in the art.

Also provided is a method of identifying an effective and/or appropriate treatment regimen for a subject with increased risk of severe lung disease, CFLD and/or MI, that includes: a) correlating the presence of one or more genetic markers described herein in a test subject or population of test subjects with severe lung disease, CFLD and/or MI for whom an effective and/or appropriate treatment regimen has been identified; and b) detecting the one or more genetic markers of step (a) in the subject, thereby identifying an effective and/or appropriate treatment regimen for the subject.

Further provided is a method of correlating one or more genetic markers described herein with an effective and/or appropriate treatment regimen for severe lung disease, CFLD, and/or MI that includes: a) detecting in a subject or a population of subjects with severe lung disease, CFLD and/or MI and for whom an effective and/or appropriate treatment regimen has been identified, the presence of one or more genetic markers described herein; and b) correlating the presence of the one or more genetic markers of step (a) with an effective treatment regimen for severe lung disease, CFLD or MI.

Examples of treatment/management regimens for severe lung disease, CFLD and MI are well known in the art. Subjects who respond well to particular treatment protocols can be analyzed for specific genetic markers and a correlation can be established according to the methods provided herein. Alternatively, subjects who respond poorly to a particular treatment regimen can also be analyzed for particular genetic markers correlated with the poor response. Then, a subject who is a candidate for treatment for severe lung disease, CFLD and/or MI can be assessed for the presence of the appropriate genetic markers and the most effective and/or appropriate treatment regimen can be provided.

In some embodiments, the methods of correlating genetic markers with treatment regimens described herein can be carried out using a computer database. Thus, in some embodiments, provided herein is a computer-assisted method of identifying a proposed therapy and/or treatment for CFLD as an effective and/or appropriate therapy and/or treatment for a subject that has CFLD, comprising the steps of: (a) storing a database of biological data for a plurality of subjects, the biological data that is being stored including for each of said plurality of subjects: (i) therapy and/or treatment type, (ii) at least one
genetic marker described herein, and (iii) at least one disease progression measure and/or symptom for severe lung disease, CFLD and/or MI from which treatment and/or therapy efficacy can be determined; and then (b) querying the database to determine the dependence on said genetic marker(s) of the effectiveness of a treatment and/or therapy type in treating and/or managing severe lung disease, CFLD, and/or MI, thereby identifying a proposed treatment and/or therapy as an effective and/or appropriate treatment and/or therapy for a subject with increased risk of severe lung disease, CFLD and/or MI.

Nonlimiting examples of disease progression measures and/or symptoms that can be monitored to determine efficacy can be determined including all of the complications and symptoms of CF, CFLD and MI as described herein and would be well known in the art.

In one embodiment, treatment information for a subject is entered into the database (through any suitable means such as a window or text interface), genetic marker information for that subject is entered into the database, and disease progression responsiveness to treatment information is entered into the database. These steps are then repeated until the desired number of subjects has been entered into the database. The database can then be queried to determine whether a particular treatment is effective for subjects carrying a particular marker or combination of markers, not effective for subjects carrying a particular marker or combination of markers, etc. Such querying can be carried out prospectively or retrospectively on the database by any suitable means, but is generally done by statistical analysis in accordance with known techniques, as described herein.

VII. Therapeutic Agents

Certain methods described herein relate to the administration of an agent that modulates the activity and/or expression of a therapeutic target described herein. Agents which may be used to modulate the expression or activity of a therapeutic target described herein include antibodies (e.g., conjugated antibodies), proteins, peptides, small molecules and inhibitory RNA molecules, e.g., siRNA molecules, shRNA, ribozymes, and antisense oligonucleotides. Such agents can be those described herein, those known in the art, or those identified through routine screening assays (e.g. the screening assays described herein).

In some embodiments, an assay is used to identify agents useful in the therapeutic methods described herein. For example, provided herein are methods of determining
whether a test compound is a candidate therapeutic agent for reducing lung disease severity, treating CFLD and/or treating MI. In general, such methods include (a) contacting a cell with the test compound and (b) detecting the expression by the cell of therapeutic target described herein (e.g. a therapeutic target associated with severe lung disease, CFLD and/or MI). A test compound that modulates the expression of a therapeutic target (for example, compared to cells treated with a placebo or untreated cells) is a candidate therapeutic agent.

Any cell can be used in the above described screening method. For example, in some embodiments the cell is a human cell. Cells used in the screen can be primary cells or a cell line. Examples of other cell lines useful in the screening assays described herein include, but are not limited to, 293-T cells, 3T3 cells, 721 cells, 9L cells, A2780 cells, A172 cells, A253 cells, A431 cells, CHO cells, COS-7 cells, HCA2 cells, HeLa cells, Jurkat cells, NIH-3T3 cells and Vera cells.

The expression of the therapeutic targets described herein can be detected using any method known in the art. For example, the expression of the therapeutic target can be detected by detecting therapeutic target mRNA using, e.g., a detectably labeled nucleic acid probe, RT-PCR, and/or microarray technology. The expression of the therapeutic target can also be detected by detecting the therapeutic target protein using, e.g., detectably labeled antibodies that have binding specificity for the therapeutic target.

In some embodiments, a cell is used in the screening assay that has been genetically engineered to facilitate the performance of the assay. For example, in some embodiments, the cell is engineered such that the therapeutic target is expressed as a heterologous protein linked to a detectable moiety (e.g. a fluorescent moiety such as GFP or a luminescent moiety such as luciferase). In other embodiments, the cell contains a nucleic acid sequence encoding a detectable moiety operably linked to the promoter of the therapeutic target. In such embodiments, rather than detecting expression of the therapeutic target, the expression of the detectable moiety is detected directly. Such cells can be generated using standard recombinant techniques well known in the art.

Agents useful in the methods of the present invention may be obtained from any available source, including systematic libraries of natural and/or synthetic compounds. Agents may also be obtained by any of the numerous approaches in combinatorial library methods known in the art, including: biological libraries; peptoid libraries (libraries of molecules having the functionalities of peptides, but with a novel, non-peptide backbone.
which are resistant to enzymatic degradation but which nevertheless remain bioactive; see, e.g., Zuckermann et al., 1994, J. Med. Chem. 37:2678-85); spatially addressable parallel solid phase or solution phase libraries; synthetic library methods requiring deconvolution; the 'one-bead one-compound' library method; and synthetic library methods using affinity chromatography selection. The biological library and peptoid library approaches are limited to peptide libraries, while the other four approaches are applicable to peptide, non-peptide oligomer or small molecule libraries of compounds (Lam, 1997, Anticancer Drug Des. 12:145).


VIII. Pharmaceutical Compositions

The agents described herein (e.g., agents that modulate the expression or activity of a therapeutic target described herein) can be incorporated into pharmaceutical compositions suitable for administration to a subject. The compositions may contain a single such agent or any combination of modulatory agents described herein and a pharmaceutically acceptable carrier. The pharmaceutical composition may further comprise additional agents useful for treating severe lung disease, CFLD, and/or MI.

As used herein, the term "pharmaceutically acceptable carrier" is intended to include any and all solvents, dispersion media, coatings, antibacterial and antifungal agents, isotonnic and absorption delaying agents, and the like, compatible with pharmaceutical administration. The use of such media and agents for pharmaceutically active substances is well known in the art. Except insofar as any conventional media or agent is incompatible with the active compound, use thereof in the compositions is contemplated. Supplementary active compounds can also be incorporated into the compositions.

A pharmaceutical composition of the invention is formulated to be compatible with its intended route of administration. Examples of routes of administration include
parenteral, intravenous, intradermal, subcutaneous, oral, transdermal (topical), transmucosal, and rectal administration.

Toxicity and therapeutic efficacy of the agents described herein can be determined by standard pharmaceutical procedures in cell cultures or experimental animals, *e.g.*, for determining the LD50 (the dose lethal to 50% of the population) and the ED50 (the dose therapeutically effective in 50% of the population). The dose ratio between toxic and therapeutic effects is the therapeutic index and it can be expressed as the ratio LD50/ED50. While compounds that exhibit toxic side effects can be used, care should be taken to design a delivery system that targets such compounds to the site of affected tissue in order to minimize potential damage to uninfected cells and, thereby, reduce side effects.

The data obtained from the cell culture assays and animal studies can be used in formulating a range of dosage for use in humans. The dosage of such compounds lies preferably within a range of circulating concentrations that include the ED50 with little or no toxicity. The dosage may vary within this range depending upon the dosage form employed and the route of administration utilized. For any compound used in the methods described herein, the therapeutically effective dose can be estimated initially from cell culture assays. A dose can be formulated in animal models to achieve a circulating plasma concentration range that includes the IC50 (*i.e.*, the concentration of the test compound which achieves a half-maximal inhibition of symptoms) as determined in cell culture. Such information can be used to more accurately determine useful doses in humans. Levels in plasma can be measured, for example, by high performance liquid chromatography.

Appropriate dosage agents depends upon a number of factors within the scope of knowledge of the ordinarily skilled physician, veterinarian, or researcher. The dose(s) of the small molecule will vary, for example, depending upon the identity, size, and condition of the subject or sample being treated, further depending upon the route by which the composition is to be administered, if applicable, and the effect which the practitioner desires the small molecule to have upon the nucleic acid or polypeptide of the invention.

**IX. Therapeutic Methods**

In some embodiments, described herein are methods for treating cystic fibrosis, reducing lung disease severity, treating and/or preventing severe lung disease, treating and/or preventing CDLD and/or treating and/or preventing MI by administering to a subject
(e.g. a subject in need thereof) an agent described herein (e.g. an agent that modulates the expression or activity of a therapeutic target described herein).

A subject in need thereof may include, for example, a subject who has or is suspected of having cystic fibrosis, a subject who lacks a wild-type CFTR gene, a subject who has a family history of CFTR and, e.g., a subject who has at least one family member that has or is suspected of having cystic fibrosis. A subject in need thereof may also be an individual having increased risk of severe lung disease, CFLD and/or MI, as determined, for example, using the methods described herein. A subject in need thereof may be a subject who carries one or more of the genetic markers described herein.

In some embodiments, the subject will be administered a pharmaceutical composition described herein. In certain embodiments the pharmaceutical composition will incorporate a therapeutic agent in an amount sufficient to deliver to a patient a therapeutically effective amount of the therapeutic agent as part of a prophylactic or therapeutic treatment. The desired concentration of the active agent will depend on absorption, inactivation, and excretion rates of the drug as well as the delivery rate of the compound. It is to be noted that dosage values may also vary with the severity of the condition to be alleviated. It is to be further understood that for any particular subject, specific dosage regimens should be adjusted over time according to the individual need and the professional judgment of the person administering or supervising the administration of the compositions. Typically, dosing will be determined using techniques known to one skilled in the art.

The dosage of the subject agent may be determined by reference to the plasma concentrations of the agent. For example, the maximum plasma concentration (Cmax) and the area under the plasma concentration-time curve from time 0 to infinity (AUC (0-4)) may be used. Dosages for the present invention include those that produce the above values for Cmax and AUC (0-4) and other dosages resulting in larger or smaller values for those parameters.

Actual dosage levels of the active ingredients in the pharmaceutical compositions of this invention may be varied so as to obtain an amount of the active ingredient which is effective to achieve the desired therapeutic response for a particular patient, composition, and mode of administration, without being toxic to the patient.
The selected dosage level will depend upon a variety of factors including the activity of the particular agent employed, the route of administration, the time of administration, the rate of excretion or metabolism of the particular compound being employed, the duration of the treatment, other drugs, compounds and/or materials used in combination with the particular compound employed, the age, sex, weight, condition, general health and prior medical history of the patient being treated, and like factors well known in the medical arts.

A physician or veterinarian having ordinary skill in the art can readily determine and prescribe the effective amount of the pharmaceutical composition required. For example, the physician or veterinarian could prescribe and/or administer doses of the agents of the invention employed in the pharmaceutical composition at levels lower than that required in order to achieve the desired therapeutic effect and gradually increase the dosage until the desired effect is achieved.

In general, a suitable daily dose of an agent described herein will be that amount of the agent which is the lowest dose effective to produce a therapeutic effect. Such an effective dose will generally depend upon the factors described above.

This invention is further illustrated by the following examples which should not be construed as limiting. The contents of all references, patents and published patent applications cited throughout this application, as well as the Figures, are incorporated herein by reference.

EXAMPLES

Example 1: Modifier loci of lung disease severity in cystic fibrosis

Genome-wide association analysis of lung disease severity in CF

A total of 3,467 CF patients are represented in three study designs (Figure 2). All patients in the GMS and 60% of the patients in the CGS and TSS populations are F508del homozygotes (denoted as F508del/F508del), while the remainder has a variety of other severe exocrine pancreatic CFTR genotypes. The three samples of CF patients showed consistent distributions of the lung disease phenotype, with the mid-range under-represented in GMS due to the extremes-of-phenotype design (Figure 3). Genotyping was
performed using the Illumina 610-Quad array® for all patients contemporaneously in a single facility under stringent quality control (see Methods Section below). Association scans for the unrelated patient designs (GMS and CGS) used an additive model adjusted for sex and with principal component correction for population substructure. Genome-wide results were combined using a directional meta-analysis approach for GMS and CGS under two different conditions for CFTR mutation status: (1) GMS and CGS, n=2,494 and (2) GMS and CGS F508del/F508del, n=1,978. The meta-analysis approach provided over 90% power to detect a SNP responsible for >2% of phenotypic variation in an additive model (Figure 4).

The combined analysis of GMS and CGS samples identified seven regions that achieved suggestive association (defined as $P \leq 1/570,725 = 1.75 \times 10^{-6}$) (Figures 5 and 6). When the analysis was restricted to patients with the same CFTR genotype (GMS and CGS F508del/F508del), the EHF-APIP region on chromosome 1pl3 had genome-wide significance at rsl2793173 ($P=3.34 \times 10^{-8}$, percentage of explained phenotype variation 1.0% in GMS, and 2.2% in CGS F508del/F508del). To further evaluate this approach, an alternative conditional likelihood approach was developed which explicitly acknowledged the selection of extremes of the lung phenotype in the GMS sample and extensive permutation analysis were performed to produce empirical p-values under the null hypothesis. Both approaches verified the genome-wide significance of the chr1 lpl3 finding (See Methods Section and Figure 7). The statistical significance of rsl2793171 was even stronger with the inclusion of several covariates shown to be related to CF lung disease ($P=9.42 \times 10^{-9}$ by meta-analysis of GMS and CGS F508del/F508del; Figure 8). The seven SNPs demonstrating association in the GMS and CGS samples were evaluated for association in the TSS sample using Merlin [5429], which accounts for family structure in a variance component framework. To be consistent with the allelic effect from the combined GMS and CGS samples, each replication test was one-sided, and the TSS patient sample for each SNP (all or F508del/F508del patients) was chosen to be consistent with the GMS and CGS patient sample set providing maximum significance. Covariates for sex and 4 principal components were included in the TSS analysis. The chromosome 11 SNP that had genome-wide significance in GMS and CGS (rsl2793 173, F508del/F508del) demonstrated significant association in the TSS sample ($P=0.006$; Bonferroni corrected $P = 0.041$ for the seven replication tests; Figure 6). Two regions of suggestive association had
less significant findings in the TSS sample: rs9268905 near HLA-DRA (P=0.032) and rs1403543 near AGTR2 (P=0.053).

A joint analysis was performed. Using all available patients, rs568529, a SNP in high LD (r2 >0.9) with rs2793173 in the EHF-APIP region, attained genome-wide significance (P=9.75 X 10-9). As in the earlier analysis, restricting to patients with the same CFTR genotype (F508del/F508del) increased the significance of the association between CF lung disease severity and the EHF-APIP region (rs568529, P=8.28 X 10-10). In the HLA class II region, a SNP (rs2395185) that is ~1kb from the suggestive SNP rs9268905 identified from GMS and CGS, approached genome-wide significance using all patients (P=9.02 X 10-8) (Figure 9). SNPs in AGTR2 remained suggestive for all patients (rs5952206, P=1.25 X10-7) and for F508del/F508del patients (rs7060450, P=3.67 X10-7).

Figure 10 shows 800kb surrounding the EHF-APIP region in detail for the genotyped SNPs. The minimum P-value was achieved in an intergenic region 3’ to both EHF and APIP. A second peak at rs286873 (P=5.62X10-7) close to EHF exhibited low linkage disequilibrium (r2<0.2) with the primary association finding (Figure 10). Multiple-SNP testing was performed to explore whether the second peak harbored other variants associated with the lung disease phenotype. After conditioning on the primary finding, rs286873 had regional statistical significance (rs2793173; regional corrected P=0.0029; Figure 11) suggesting that several genetic variants may be responsible for association in the EHF-APIP region. Finally, MACH imputation was used to determine if any other variants not typed by the Illumina SNP array associated with lung function. Analysis of the imputed SNPs in the region identified the same EHF/APIP interval, with minimum P=1.45X10-8, at rs535719, at a position 19kb closer to APIP than the genotyped SNP rs12793173. As an adjunct to testing of SNP genotype association, testing of association of the Consortium lung phenotype was performed with copy number variation, for models using estimates of both total copy number and allele-specific copy number (Online Methods). None of these models met genome-wide significance for the regions with variant frequency >1% (illustrative Manhattan plot in Figure 12).

Linkage of lung disease severity in CF to chromosome 20q13.2

Linkage analysis of 486 sibling pairs using 19,566 SNPs selected from the Illumina array revealed a genome-wide significant multipoint LOD score of 5.033 at rs481 1626,
located at 53.81 Mb (~85cM) on chromosome 20q13.2 (nominal P = 7.9 X 10^-7; genome-wide P=2.3 X 10^-3). Another noteworthy linkage signal was on chromosome lp22.21, with multipoint LOD score of 2.48 for rs94103 at 91.07 Mb (119 cM). As body mass index is an important covariate of CF lung function (Figure 8) and measures of human obesity have been linked to this region of chromosome 20, linkage analysis was performed using an average body mass index Z score (BMI-Z) as a covariate. Inclusion of BMI-Z increased the LOD score for the linkage peak on 20q13.2 (Figure 13; the highest multipoint was for rs481 1645 (LOD=5.72; genome-wide P=5.05e-04), which is 0.07cM (0.13Mb) from and immediately adjacent to rs481 1626 on the linkage panel) and revealed suggestive linkage (rs1990398; LOD 2.21) on chromosome Xp21 at 27.41 Mb (43 cM). In contrast, the linkage peak on chromosome lp22.21 decreased to LOD 1.67 when BMI was used as a covariate. Using three different methods, it was determined that the QTL on chromosome 20 may account for almost 50% of the variation in lung function in the CF sibling pairs (Figure 14), however this estimate is likely to be biased upward due to winner's curse.

A region of 1.314 Mb on 20q13.2, demarcated by 1 LOD unit below the maximum (when BMI-Z was used as a covariate), was selected for further analysis to determine if any SNPs within the chromosome 20 linkage peak demonstrated association with lung function in the combined GMS and CGS samples. In the center of the region (encompassing 307 SNPs, rs2225092 to rs6999), a cluster of SNPs 3’ of CBLN4 generated the lowest P-values in the combined GMS and CGS F508del/F508del samples (Figure 15, lower panel). These SNPs (rs6092179, rs6024437, rs8125625, rs6024454 and rs6024460) displayed high LD with each other (r2 > 0.8) and were located in a 30kb region (53.79 Mb to 53.82 Mb). The SNP with the lowest P-value (rs6024460; P=1.34 X 10^-4) was significant for the region (Bonferroni corrected P =0.041). Importantly, the cluster of SNPs exhibiting allelic association with lung function in the unrelated patient samples are found in the same LD block as the SNPs with the highest LOD scores in the family-based sample (see above).

A combined false discovery rate approach corroborates genome-wide significance of loci on chromosomes 11 and 20

To evaluate association and linkage results in a single genome-wide framework, a combined analysis was performed that integrated the linkage data from TSS into the association data from GMS and CGS F508del/F508del. In essence, linkage information
was used to reprioritize genome-wide association results using extensions of the false
discovery rate (FDR) control methodology via the stratified FDR (SFDR) and weighted
FDR (WFDR). The linkage-weighted q-values (genome-wide adjusted P-values that
control FDR) reflective of the combined evidence for a gene modifier at a given SNP,
within the GWAS design were obtained, and GWAS results after accounting for the linkage
evidence were re-ordered. Presented herein are data from the WFDR; all results were
confirmed using the SFDR. SNPs with q-values less than 0.05 were declared to be
genome-wide significant (Figure 16). SNPs in the EHF-APIP region on chromosome 11
are highly significant with or without the linkage information, because of the initial strong
evidence for association. In contrast, the q-values for SNPs under the linkage peak on
chromosome 20 change considerably after accounting for the linkage signal. The results
presented in Figure 16 illustrate that the linked SNPs on chromosome 20 are now top
ranked genome-wide, while they were ranked 154th or lower by the GWAS evidence alone,
prior to incorporating the linkage information. A rank-based Manhattan plot of the
combined results demonstrates that chromosome 11 and chromosome 20 both provide
genome-wide significant results (Figure 17).

Methods

Recruitment. The TSS sample consisted of 486 affected sibling pairs (904 CF
patients: 420 families with 2 siblings deriving 420 pairs, 20 families with 3 siblings
deriving 60 pairs & 1 family with 4 siblings deriving 6 sibling pairs) recruited by the TSS.
An additional 69 singletons from the TSS study were included for association analysis. All
TSS patients and families were recruited based on having a surviving affected sibling as
previously described. Written informed consent was obtained from all patients over 18
years of age. Parental or guardian consent was obtained for patients less than 18 years old
along with assent from patients between 6 and 17 years old. Studies were approved by the
Institutional Review Boards of Johns Hopkins University, UNC, CW and the Research
Institute at the Hospital for Sick Children, Toronto, Canada.

Lung disease severity phenotyping. In CF, FEV1 is recognized as producing the
most clinically useful measurements of lung function and a known predictor of survival.
However, comparison of disease severity by FEV1 across a broad range of ages is
confounded by the decline in FEV1 over time in CF patients, and by mortality attrition. In
brief, age-specific CF percentile values of FEV1 were calculated for each patient using 3 years of data in patients 6 year or older, using the Kulich-derived U.S. (national) CF percentiles (relative to other CF patients of the same age, sex, and height, and then adjusted for CF age-specific mortality. The resulting quantitative phenotypes were distributed as expected based on ascertainment (Figure 3). This quantitative phenotype was also highly correlated with the Schluchter survival phenotype (r² = 0.91), when compared for the GMS patients.

Additional phenotyping. All patients carried severe CFTR mutations on both alleles that are known to confer pancreatic insufficiency. Sex was based on self-report and consistent genotype. Average BMI Z-score, used to stratify patients by nutritional status, was derived from the body mass index (kg/m²) calculated from height and weight measurements over the same time period (3 years duration) used to calculate the lung function phenotype. Standard deviation scores (Z-scores) were then calculated using CDC reference equations. After removal of 0.2% of data points due to inconsistent height or weight values, the resulting values were averaged to produce the BMI-Z covariate.

Genotyping and quality control. DNA derived from either whole blood or transformed lymphocyte cell lines was hybridized to the Illumina 610-Quad genotyping platform at Genome Quebec facilities (McGill University and Genome Quebec Innovation Centre, Montreal) using the 96-well plate format. The plates containing the DNA samples were loaded at the respective lead institutions with a balance of sex and lung severity. Two CEPH DNA controls and one randomly chosen replicate sample were included per plate for quality control. Illumina BeadStudio software was used to call genotype. Sample identity was confirmed by comparing SNP calls to a Sequenom fingerprinting panel. Any discrepancies were resolved or rerun. Further quality control for SNPs and samples was conducted as summarized below.

The quality of the SNP calls was judged to be very high, with the discordance rate between duplicate samples calculated at 0.004% in GMS, and similar for the other studies. SNPs monomorphic across the studies were removed from analysis. SNPs were also removed if they showed a missing data rate > 10%. Hardy-Weinberg testing was not used as an initial SNP filter, to allow discovery of true associations that might exhibit departures from equilibrium. The trios (mother, father, child) within TSS offered the opportunity to estimate SNP call error rates, and missing data rates. Using these trios, error rates from
homozygote to heterozygote, homozygote to other homozygote, and heterozygote to homozygote were calculated, and SNPs with >1% Mendelian error rates were removed from analyses in all studies. Finally, 570,725 SNPs from autosomes and the X chromosome were selected for analyses, as well as 158 SNPs on chromosome Y and 138 mitochondrial SNPs.

Patient samples were excluded if the call rate in an initial screen was below 98%. Identity by state and identity by descent inferences were used to identify unexpected relatives in the datasets and samples with duplicate enrollment or plated in more than one study. Sex was confirmed by counting heterozygous X-chromosome genotypes and the number of called Y-chromosome genotypes. Unresolved sex mismatches and aneuploid patients were excluded. Samples were also examined to ensure no sample had an excess or deficit of heterozygous SNP calls more than 5 standard deviations from the mean heterozygosity of 31.6%. Additionally, quality was also assessed using a subset of the UNC cohort (917 patients) that had been previously genotyped for 1536 SNPs in candidate genes using Illumina GoldenGate technology. Of these SNPs, 542 were also found to give high quality SNP calls on the Illumina 610-Quad. Discordance between the genotype calls across the two platforms was 0.07%. For family-based samples, Mendelian consistency was checked for all trios. Samples and families with more than 5% Mendelian errors were excluded. In all, 28 (GMS6; CGS 17, TSS 5) patient samples were excluded from analysis due to genotyping failure or apparent artifacts, 2 GMS samples were excluded due to outlying ancestry (as evidenced by PC analysis) and 8 GMS samples were excluded for excessive (> second degree relation) of identify-by-descent proportions with other samples in the study. All of the reported significant and suggestive loci were subjected to additional scrutiny using Illumina GenomeStudio VI. 0.2 genotyping module VI. 0.10. with the GenTrain clustering algorithm and manually-assisted genotyping to ensure high-quality calling and minimal potential impact of copy-number variation.

Genome-wide association testing. Regression analyses for the common lung phenotype were performed separately for the three study samples GMS, all CGS patients, and the CGS F508del/F508del patients. PLINK v. 1.07 (Am J Hum Genet. 2007 Sep;81(3):559-75) was used for each analysis, using an additive genetic model while adjusting for sex and genotype-derived principal components (PCs).
For the GMS and CGS samples, PCs were obtained from SMARTPCA using a thinned set of 70,000 markers, derived separately for each study sample/subset, as described in Li et al. (Clin Genet 2010). Eigenvalue analysis resulted in the choice of 4 PCs for the relatively homogeneous GMS sample, and 7 PCs for each of the CGS study samples. To acknowledge potentially differing standard errors across the sample designs due to the extremes-of-phenotype GMS design, a standard weighted meta-analysis z-statistic (Houwelingen et al, Statist. Med. 2002; 21:589-624) was constructed as a combined association statistic for primary GWAS analyses (GMS and CGS, GMS and CGS F508del/F508del). Using each of the genetic effect z-statistics for the GMS and CGS samples, the combined statistic was 

\[ z = w_{GMS}^GMS z_{GMS} + w_{CGS}^CGS z_{CGS} \]

with weights inversely proportional to the standard errors, and a common reference allele to ensure directional consistency of risk effects. Suggestive association was declared for P-values lower than the approximate threshold \( l/(\text{number of SNPs}) = l/570, 725 = 1.75 \times 10^{-6} \). Significant association was declared using the conservative Bonferroni threshold \( P < 0.05/570, 725 = 8.76 \times 10^{-8} \).

Slight variation in the number of informative markers for various analyses was of no consequence in declaring significance. Similarly, genotypes among males of X-chromosome SNPs were encoded according to default PLINK settings (0 or 1 copies of the minor allele), but an alternative coding to 0 vs. 2 copies resulted in no qualitative changes in conclusions or in identification of the most significant SNPs on chromosome X.

To further investigate and refine the significance threshold, the entire association testing procedures for GMS and CGS, GMS and CGS F508del/F508del were performed for each of 1,000 permutations, with genotypes permuted relative to the phenotypes and covariates. Although PCs are derived from genotypes, they represent patient-specific ancestry, and thus remained aligned with the phenotype. From this pool of study sample permutations, results were randomly drawn to obtain 10,000 null permutations for each of the meta-analyses. The obtained significance thresholds for a genome-wide error rate of 0.05 were \( P = 1.07 \times 10^{-7} \) (GMS and CGS) and \( P = 1.05 \times 10^{-7} \) (GMS and CGS F508del/F508del), illustrating the conservativeness of the Bonferroni threshold. Consequently, for either analysis, a P-value of 5X10-8 is sufficient to achieve false positive error control at a genome-wide value of 0.05, even considering the multiple comparisons implied by two separate GWAS analyses.
Association analysis for SNPs in TSS was performed in 1,042 patients using regression as implemented in the variance components framework in Merlin9. Missing genotypes (0.125%) were inferred by Merlin to optimize the power of associations. Additive model regression was conditioned on linkage and covariates for relatedness of patients, sex, and 4 principal components to control for population stratification. Principal component analysis was performed as described (Li et al. Clin Genet 2010). Association analysis was performed on the 557 TSS F508del/F508del patients in the same manner. Joint analyses of the GMS, CGS and TSS associations proceeded with the meta-analysis approach as described above, with the three studies contributing to the weighted direction-consistent z-statistic.

**Combined conditional likelihood approach to association testing.** The testing approach described above preserves false positive error control, but it was reasoned that additional power might be achieved by explicit acknowledgement of the GMS sampling design. A case-control approach would artificially dichotomize the data, thus losing power due to variation of the phenotype within the extremes group. An efficient conditional likelihood method for handling extremes of phenotype association data has been described (Huang and Lin, Am J Hum Genet. 2007 March; 80(3): 567-576), but this approach requires sampling within a predefined region of the phenotype (e.g. precise tails). However, use of this approach was not possible for the data, as the GMS study employed initial entry criteria which had only an approximate (though strong) correspondence to the Consortium lung phenotype, and the lung function measurements were further refined by retrospective record-based evaluation and subsequent follow-up. To fully exploit these data, a novel but straightforward approach was devised that appropriately conditions on the GMS sampling scheme. The approach uses the assumption that the CGS dataset represents a random population sample, whereas the contribution of the GMS dataset is conditional on the observed phenotypes, where the phenotype selection criteria are completely unspecified. The SNP genotypes g were recorded as the number of minor alleles at the locus, and common lung phenotypes y in each study were pre-adjusted for sex and the study-specific PCs described above. A population additive association model

\[ y = \beta_0 + \beta_g + \epsilon, \epsilon \sim N(0, \sigma^2) \]  

was assumed. The full likelihood conditioned on the GMS
phenotype sampling was
\[ L = p(g_{CGS} > y_{CGS}) p(y_{GMS} | g_{GMS}) \]
\[ = p(g_{CGS} > y_{CGS}) p(y_{GMS} | g_{GMS}) = p(g_{CGS} > y_{CGS}) p(y_{GMS} | g_{GMS}) \]
where \( p(y_{GMS} | \beta_0, \beta_1, \sigma^2) = \sum_{j=0}^{2} p(g_{GMS} = j) p(y_{GMS} | g_{GMS} = j, \beta_0, \beta_1, \sigma^2) \). Finally, for each SNP the statistic \( 2 (\log\text{-likelihood ratio)}, \) with the null likelihood assuming \( \beta_1 = 0 \).

P-values were obtained by comparison of the log-likelihood ratio statistic to \( \chi^2 \) was computed. The approach assumes that the effect sizes are the same in GMS and CGS, which is true under the null hypothesis.

**Power Analyses.** Power analyses for the combined GMS and CGS studies were performed by assuming an underlying additive allelic genetic model, with assumed effect \( \beta_1 \) on the average phenotype for each additional copy of the minor allele (as described in the additive model immediately above). Only the results for GMS + CGS F508del/F508del are shown in Figure 4. The power derived from all patients is higher due to the larger sample size. Using the CGS sample to estimate the population mean and variance and assuming the underlying effect size, the posterior probabilities of the genotypes were computed using the actual observed phenotypes (for both GMS and CGS), and they were used to simulate accompanying genotypes. In this manner, the simulated data respected the GMS extremes-of-phenotype design and any exaggeration of apparent effect sizes due to the design. Then, for each simulation, the weighted meta-analysis procedure was performed, with the meta-analysis P-value compared to the genome-wide threshold 5X10^-8.

**SNP genotype imputation.** Using the hidden Markov model algorithm implemented in MACH (available online at sph.umich.edu/csg/abecasis/mach/) and IMPUTE (available online at /mathgen.stats.ox.ac.uk/impute/impute.html), genotype imputation was conducted for 1162 patients recruited by the University of North Carolina site and 1254 1,254 self-reported "Caucasian" patients recruited by the Toronto site. As some these individuals were later used for the TSS study, association analyses were performed only for the unique subsets in GMS and CGS, respectively, as give in Figure 2. The reference sample for imputation was the 60 CEU samples from HapMap Phases I and II. For each subject, the genotype for each SNP was reported as a dosage value (a continuous value between 0 and 2), reflecting the expected number of copies of a reference allele at that SNP, conditional on the directly observed genotypes and the phased CEU haplotypes. MACH was used to
impute autosomal SNPs and IMPUTE for chromosome X SNPs. Imputation yielded genotype data for -2,544,000 autosomal and -65,000 chromosome X SNPs, respectively. Approximately 36,000 SNPs with estimated imputation accuracy less than 0.3 (using MACH's $r^2$ accuracy measure) were discarded. Refined genotype imputation was conducted for GMS and CGS SNPs on chromosomes 11 and 20, using HapMap Phase III data, thereby increasing the number of novel imputed SNPs by -22,000 across the two chromosomes.

*Copy-number analysis.* Copy number variations (CNVs) were detected using both pennCNV (2008Nov19 version) and genoCNV (version 1.08) using default parameters. Samples with lower quality were dropped, which were initially identified by relatively larger number of copy number calls and were confirmed by visual inspection. In total 1103 and 1303 samples were used for CNV association in GMS and CGS, respectively. CNVs harboring fewer than 5 probes were filtered out and only the probes with copy number changes in ≥ 1% of the samples were used in the following association studies, which results in 3,008/4,868 probes from genoCNV/pennCNV in GMS study, and 3015/4663 probes for genoCNV /pennCNV in CGS.

In addition to overall copy number, genoCNV can also identify allele specific copy number. Principal components (PCs) identified from the genotype data were used to account for population stratification, as described for the association study methods. Two models were used to evaluate the association between trait and copy number call, probe by probe. For the $j$-th probe, the models (i) \( \text{trait} = \text{sex} + \text{PCs} + c_n\) ; (ii) \( \text{trait} = \text{sex} + \text{PCs} + c_n + c_n^*\text{sex} \), were considered, where \( c_n \) indicates the total copy number of the \( j \)-th probe. In addition, genoCNV is able to identify allele-specific copy number. Let \( a_{c_j} \) (allele-specific copy number contrast) be the number of \( B \) alleles minus the number of \( A \) alleles at the \( j \)-th probe. Two additional models were considered: (iii) \( \text{trait} = \text{sex} + \text{PCs} + c_n + a_{c_j} \); (iv) \( \text{trait} = \text{sex} + \text{PCs} + c_n + a_{c_j} + c_n^*\text{sex} + a_{c_j}^*\text{sex} \). In addition to overall copy number, genoCNV can also identify allele-specific copy number.

*Linkage Marker Selection.* 19,566 SNPs were selected from the Illumina 610k SNP array based on minor allele frequency >0.4 and $r^2$<0.01 between adjacent SNPs using Merlin. HapMap 2 (www.hapmap.org) recombination data were used to interpolate physical position (in basepairs) with genetic map (in centimorgans). Average distance between markers was 0.18 cM or 0.13 Mbp. Whenever a physical position did not have
match in HapMap, genetic position was estimated assuming uniform recombination rates between adjacent SNPs with established genetic positions. The average information content of the markers was -0.9 (multipoint) and -0.31 (two-point) throughout the genome as determined by Merlin.

Linkage Analysis. Linkage analysis for the lung phenotype was performed using the variance components method implemented in Merlin (Multipoint Engine for Rapid Likelihood Inference). Linkage was also performed using SOLAR (Sequential Oligogenic Linkage Analysis Routines). Multipoint IBD probabilities generated by Merlin were used for both linkage programs. Very similar results were obtained when linkage analysis was performed using Merlin or SOLAR. Covariates for linkage were sex or sex and average BMI Z-score. Two-point and multi-point linkage maps were generated with and without covariates. Multipoint logarithm of the odds (LOD) of linkage >2.0 was considered suggestive and LOD>3.7 was considered to be of genome-wide significance.

WFDR and SFDR methods. Let $P_i$ be the p-value of an association test for SNP $i$, $i = 1, \ldots, m$. FDR control can be achieved by converting the p-values to the corresponding q-values. SNPs with q-values less than the FDR threshold value (e.g. $\gamma = 0.05$) are declared significant. The expected proportion of false positives among all the positives is then controlled at the $\gamma$ level. Note that although q-value differs from p-value, there is a monotonic relationship between the two. Thus ranking SNPs by p-value or q-value are equivalent.

Let $Z_i$ be the linkage score of SNP $i$ obtained from a previous GWL study using either allele-sharing or parametric approaches. For the SFDR method, $m$ SNPs are divided into $K$ disjoint strata based on the prior linkage information. Without loss of generality, consider $K = 2$ and assign each SNP $i$ to stratum 1 (the high priority group) or stratum 2 (the low priority group) according to whether the linkage score $Z_i$ exceeds a threshold $C$ ($C=3.3$ was used corresponding to significant linkage discussed by Lander and Kruglyak, 1995). FDR control is then applied separately in each stratum at the same $\gamma$ level (Sun et al., 2006), i.e., q-values are calculated separately for each stratum of SNPs. Ranks of the GWAS SNPs are determined by the corresponding q-values and the original association p-values are used to break any ties among the q-values.

In contrast, WFDR calculates a weighting factor $W_i$ for each SNP $i$ with weights subject to two constraints: $W_i \geq 0$ and $\bar{W} = \sum_i W_i / m = 1$. The weight $W_i$ is proportional
to the linkage signal $Z_i$ for SNP $i$ (e.g. $W_i = \exp(B \cdot Z_i) / \nu$, $\nu = \sum_i \exp(i \cdot Z^\text{lm} \cdot \beta, \beta$, and $B = \gamma$) (Roeder et al., 2006), and the FDR procedure is applied to the set of weight-adjusted p-values, $P/W_i = \ldots, m$. $B = 2$ was used in the present analysis. The WFDR and SFDR were implemented in a perl program called SFDR, available at http://www.utstat.toronto.edu/sun/Software/SFDR/index.html.

Example 2: Modifier loci of cystic fibrosis related liver disease

Genetic Variation that Modifies Clinical Phenotype in CF (Liver Disease)

As described herein, a GWAS study of 294 CF patients with CFLD and 1,837 CF patients without CFLD was used to identify a genetic locus on chr 21q22.3 that likely causes severe CF liver disease (CFLD) through non-CFTR genetic variation. The 6 SNPs in the 3’ end of SLC19A1 (rs914232; rs2330183; rs2838956; rs051266; rs4819130; 30 rs3788190) and 4 SNPs in the 3’ end of COL18A1/endostatin (rs2236483; rs2838950; rs3753019; rs2483377) strongly associate with CFLD.

Mechanism of liver disease pathogenesis in CF.

Only 5-7% of CF patients develop severe liver disease, even though most CF patients have abnormal hepatic biochemical markers, changes in tissue architecture on biopsy, or evidence of HSC activation. In CF, signaling molecules are released from activated cholangiocytes, in response to abnormal (sluggish) bile flow, which initiates stellate cell activation and proliferation, and leads to increased collagen production and fibrosis (Figure 18). Since most CF patients have some evidence of liver disease, there appears to be a state of susceptibility that leads to more significant cirrhosis if additional stimuli are operative, which is congruent with the recent report that the Z-allele is a strong risk factor (OR ~5) for the development of CFLD (overall $p \sim 9 \times 10^{-9}$).

Performing a GWAS.

A GWAS in nearly 300 CFLD patients was pursued. This was done in conjunction with a GWAS in nearly 4,200 CF patients, who were being studied for gene modifiers of CF lung disease by the North American CF Gene Modifier Consortium. The results from
this GWAS also provided a special opportunity for CFLD, because there were ample non-CFLD "control" CF patients among those studied for lung disease.

The genotyping was performed on an Illumina 610-Quad platform at Genome Quebec facilities. Extensive measures were taken to ensure quality by adding replicate samples to each plate. In addition, there were extensive data cleaning methods that were undertaken by 2 independent investigators, as well as reclustering and manual analysis of selected SNPs. After data cleaning, 570,725 SNPs from autosomes and the X chromosome were approved for analysis. The patient samples were also scrutinized carefully, and identical by state (IBS) comparisons were used to exclude unexpected duplicate samples or related individuals.

GWAS results.

The overall results for the GWAS are shown in a "Manhattan" plot of 294 CFLD patients (184 males and 110 females) vs. 1,837 non-CFLD "control" CF patients (> 15 years old) (Figure 19), based on p-values calculated using a logistic regression model (run with PLINK), and using covariates of gender, principal components (7 "PCs") and a binary variable indicating whether a patient is homozygous for L1F508 (similar results were obtained when co-variates included whether patient had either one or two copies of L1F508). There are SNPs on Chr. 21 that are near genome-wide significance (p::: :: 1 x 10^-7) for association with CFLD, and the genes in this region (SLC19A1/COL18A1) are discussed herein. There are also SNPs on Chr 1, 16, and 17 that are near or above the "suggestive" threshold for association. The region of Chr 1 is driven by association among males (discussed herein). The SNPs on Chr 16 are near DNAH3 and TMEM159. The latter gene (TMEM159) is also known as promethin, which is one of the genes upregulated in a mouse model of hepatic steatosis. The SNP on Chr 17 (RPA1) codes for a DNA repair protein.

Chr 21 region.

A closer (blown-up) view of the Chr. 21 region (Figure 20) shows that the SNPs with the strongest association with CFLD are in the 3’ end of two genes: SLC19A1 and COL18A1. SLC19A1 is a folate transporter without a commonly recognized role in liver fibrosis, although moderately high folic acid supplementation exacerbates the development of fibrosis in rats with experimentally CC14-induced liver fibrosis. Rats that received the
folate supplement had 1) higher collagen content, 2) more activated stellate cells, and 3) more apoptotic hepatocytes in liver tissue, compared to controls. In addition, the supplemented rats had alterations in expression of genes related to apoptosis and fibrosis. Thus the highest ranked SNPs in this region exhibited a significant association with differential expression of \textit{SLC19A1} \((p \sim 10^{-5}), \text{i.e., an "eQTL," in -460 samples of LCIs from CF patients. Further, the "risk allele" for these SNPs was associated with increased expression of \textit{SLC19A1}, in line with the above experimental findings. Therefore, it is likely that genetic variation of \textit{SLC19A1} and/or variation in gene expression of \textit{SLC19A1} contributes to the development of CFLD, \text{i.e., biliary cirrhosis and portal hypertension.}

\textit{COL18A1} is a basement membrane collagen that is predominantly expressed in highly vascularized organs, such as liver and lung. Whereas no association was seen of SNPs near \textit{COL18A1} with lung disease severity, there is a very strong association with CFLD (Figure 20). There are several reasons that \textit{COL18A1} is a strong, biologically plausible candidate to modify fibrosis in the CF liver. First, \textit{COL18A1} is expressed (produced) mainly in hepatocytes and activated cholangiocytes and bile duct epithelia of normal and fibrotic livers (Schuppan D, \textit{Lancet} 1998), which is in striking contrast to the other extracellular matrix proteins, including collagens, which are mainly produced by hepatic stellate cells and myofibroblasts. Second, hepatocyte \textit{COL18A1} expression remains constant in acute fibrogenesis (in contrast to greatly upregulated pro collagen a1 and TIMP1), while it is upregulated in regions of biliary fibrosis and cirrhosis (Schuppan D, \textit{Lancet} 1998). Third, \textit{COL18A1}, which carries structural properties of both a collagen and a proteoglycan, is upregulated during angiogenesis and early extracellular matrix remodeling, and \textit{COL18A1} has a role in binding other basement membrane proteins and proteoglycans, as well as cell surface receptors, playing a role in basement membrane reorganization and angiogenesis (Marneros AG, Olson BR, FASEB J 2005). Finally, the C-terminal domain of \textit{COL18A1} is proteolytically cleaved to release a 20K peptide, endostatin, which exhibits potent antiangiogenic effects and for which functionally relevant genetic polymorphisms exist.

Endostatin induces apoptosis in endothelial cells, and has potent antiangiogenic activity. The liver's response to injury involves angiogenesis sinusoidal remodeling, and pericyte (i.e., HSP) expansion. Thus, genes related to angiogenesis may be important modifiers of liver fibrogenesis, including \textit{COL18A1/endostatin}. Animal experiments
suggest that anti-angiogenic agents (such as VEGF inhibitors) might provide an antifibrotic approach, but complete inhibition of angiogenesis might limit hepatic blood flow, with adverse consequences, especially in biliary fibrosis (Patsenker E, Hepatology 2009). There is a delicate balance between angiogenesis vs. anti-angiogenesis in response to liver injury/repair, while the role of COL18A1/endostatin in hepatic fibrosis remains to be examined. Loss of function mutations in COL18A1 in humans and experimental animals lead to vitreoretinal degeneration and hydrocephalus, but no overt liver disease; but deletion of the COL18A1 gene leads to enhanced arterial and cardiac angiogenesis, and delayed dermal wound healing (Li Q and Olsen BR, Am J Pathol 2004; Moulton KS, Circulation 2004; Seppinen L, Matrix Bio/2008). The pathophysiology in CFLD could involve both the loss of CFTR function and variant function of COL18A1 and/or endostatin. The apoptotic (proliferative) effects of endostatin (collagen XVIII) are thought to be selective for endothelial cells, but its (and intact collagen XVIII’s) effects on apoptosis (and proliferation) of hepatocytes, cholangiocytes, and hepatic stellate cells merits further investigation in hepatocytes and cholangiocytes.

**Example 3: Modifier loci of meconium ileus**

**Determination of Gene Modifiers of MI**

*Genotyping and Quality Control*

As part of the North American CF Consortium, all CF patients at Genome Quebec (McGill University and Genome Quebec Innovation Centre, Montreal) have been genotyped. Illumina BeadStudio software was used to call the SNP genotypes across the entire set and quality control conducted concurrently. After quality control, samples were analyzed for lung function modifiers. Subsequently, the samples from each site were pooled for a meta-analysis of lung function. A total of 3655 Caucasian patients (ethnicity determined via principal component analysis), analyzed at 556,445 autosomal SNPs and 14,280 chromosome X SNPs, were included in MI modifier analysis; **Figure 21** summarizes the consortium samples used in the MI GWAS, by demographics. Cryptic relatedness across all study samples was determined using identity by decent statistics calculated using PLINK and PREST-plus, with five first-degree relations previously unknown, identified.
GWAS Analysis and Genome-wide Significant Results

Given the presence of the subset of related samples from the TSS collection and the identified cryptic relatedness, Generalized Estimating Equations (GEE) was used with an exchangeable correlation matrix to assess the evidence for association between MI and each SNP in the total sample of n=3655 CF patients, adjusting for ascertainment site. Genotypes of the autosomal SNPs were coded additively as 0, 1 and 2 minor alleles, and the X chromosome SNPs were coded 0, 1 and 2 for females and 0 and 2 for males. In parallel, the subset of unrelated samples (n=3055) was also analyzed using logistic regression, adjusting for site and 7 principal components estimated using eigenstrat to assess the evidence for population stratification. Adjustment for principal components had little effect on the results and was deemed unnecessary. The two analyses, logistic regression with PCs and GEE without PCs provided very similar association results with no evidence of stratification.

Figure 22 provides a QQ-plot of the GEE analysis (n=3655) indicating significant departure from random (or chance) findings. The most significant set of SNPs are located on the X chromosome, in the promoter region of SLC6A14, a gene known to encode a protein at the intestinal brush border. To interpret association results from the X chromosome, multiple factors such as how male and female genotypes were coded for analysis, X-inactivation status, sex-specific odds ratios (OR), and Hardy-Weinberg equilibrium in females are also considered. rs3788766 (minor allele frequency (MAF) = 0.41) provided the smallest p-value (2.09x10^{-12}), exceeding the strict genome-wide significant threshold (p<5x10^{-8}). The rs3788766 normalized intensity plot was as expected for an X chromosome SNP. When analyzed separately by males (p=3.3x10^{-9}) and females (p=1.4x10^{-6}, calculated from an additive model) both sexes contribute association evidence, all three ascertainment sites contribute to the association, and a single C allele is sufficient to decrease one's MI risk, under the assumption of X-inactivation; OR=0.44 (95%CI 0.29-0.66) in males and OR(CC vs TT)=0.52 (95%CI 0.34-0.78) and OR(CT vs TT)=0.59 (95%CI 0.45-0.78) in females (Figure 23).

The regional plot (Figure 24) suggests that all SNPs associated in SLC6A14 are in linkage disequilibrium (LD) with rs3788766 (measured by r^2). Conditional analysis confirmed the absence of allelic heterogeneity. CNVs in the SLC6A14 region have not been reported in the database of genomic variants (see projects.tcag.ca/variation/).
Adjustment for CFTR mutation, genome-wide, had little effect on the results indicating that the genotypic associations are not confounded by CFTR mutation, including those that are known to lead to appropriately localized, but dysfunctional CFTR protein (e.g. CFTRG551D) as well as misfolded CFTRAF508 that is retained in the ER and rapidly degraded. However, due to low frequencies of patients carrying mutations other than AF508, there is little power to detect interactions with CFTR.

**GWAS-HD and Suggestive Association Results**

Although SLC6A14 contains SNPs that reach a strict genome-wide significance criterion, these particular SNPs cannot explain all the MI genotypic variance (pseudo R²=0.027), and there are likely other genes which contribute to MI. The GWAS results indicate a number of SNPs that provide suggestive association evidence (p<10⁻⁵) (Figure 25a - using all 3655 CF patients, and Figure 25b - restricted to 2675 AF508 homozygous CF patients), including those in SLC26A9, a gene member of the SLC 26 family that encodes a protein in the apical plasma membrane of the gut lining and other organs. MAF and HWE p-value were calculated based on 2505 independent controls (no-MI). The p-value was obtained from association analyses using all available samples adjusting for a center/site covariate, via the GEE model that accounts for the correlation between a subset of samples that are related. The q-value is genome-wide adjusted p-value that controls the False Discovery Rate. The rank is the rank of a SNP at the genome-wide level based on its association p-value alone (GWAS) or combined association evidence incorporating the apical plasma membrane hypothesis (GWAS-HD). **Figure 25c** provides OR estimates for all SNPs in **Figure 25a** with a q-value < 0.05, as well as identifying the minor allele.

Given that loss of CFTR function leads to perturbed epithelial transport, it was considered that coincident phenotypes including MI reflect residual or adapted transport capability. Given the distinct localization of many transport-relevant proteins in polarized epithelial tissue, it was hypothesized that constituents of the apical plasma membrane, where CFTR normally resides, may be contributors to the MI phenotype. To test this hypothesis in the MI GWAS data, the **Hypothesis-driven GWAS (GWAS-HD)** methodology was developed and applied by: 1) prioritizing the genome by generating a list of genes that encode proteins that localize to the apical plasma membrane (as set out in **Figure 26**) - this list of 157 genes was annotated by AmiGO version 1.7 (downloaded March 28th, 2010),
based on GO consortium, generated by location search phrase *apical plasma membrane* with restriction to *homo sapiens*, and assigning GWAS SNPs to a high priority group (SNPs within the boundaries of these genes) or a low priority group (all others); 2) using the stratified FDR approach (SFDR) to weight GWAS p-values and determine genome-wide significance for given loci; and 3) performing permutation testing to determine the statistical significance of the hypothesis.

The apical plasma membrane list consisted of 151 genes spanning 3,723 GWAS SNPs, although eight genes were not tagged by any of the ~550k GWAS SNPs. SNPs were assigned to genes if they were within ±10 kb of the gene boundaries as annotated from public databases. Although CFTR and many solute transporters are included, SLC6A14 is not on this gene list despite it being on the apical brush border membrane, likely reflecting the high specialization of this type of intestinal cavity and a limitation of the GO annotation that was accepted without additional curation.

**Figure 27a** provides the qq-plot of the 3,723 SNPs in the apical membrane list, whereas Figure 3B provides the qq-plot of all of the remaining SNPs. Note that in this first analysis, X chromosome SNPs were omitted to exclude SNPs reaching genome-wide significance. Clearly there is substantial deviation in the observed p-values from expected in **Figure 27a** in contrast to **Figure 27b**. The deviation appears to begin early in **Figure 27a**, at approximately a chi-square value of 5 (corresponding to p-value of 0.025), and this deviation consists of 175 SNPs spanning 40 different genes from the original list of 143 genes. However, since independence is one of the underlying assumptions in this plot, it is not clear that the deviation reflected in the figure is significant and that there is truly a preponderance of apical plasma membrane constituents that can contribute to MI susceptibility. As part of GWAS-HD this departure was formally tested (accounting for LD correlation between SNPs in a gene), as well as determined the significance of individual SNPs after prioritization.

To test the hypothesis, the MI case control phenotype status was permuted 1000 times to obtain 1000 null simulated datasets that retain the same LD pattern of markers as in the apical plasma membrane gene list. The association evidence was then reanalyzed across the 143 genes of interest for each of the 1000 permuted replicates. The qq-plot in **Figure 28a** displays the distribution of p-values across these 1000 null datasets (black curves). The red curve is the observed result, which is clearly an extreme as compared to
the null distribution of black lines (p<0.001 or 0 out of 1000 permuted datasets are more
extreme than the observed one, calculated using a Shapiro-Wilks test); this implies that
multiple genes on the apical plasma membrane are associated with MI. For comparison, a
null hypothesis/list of membrane-localized genes was constructed for which no relationship
with MI was expected. This list consisted of all gene products present in the nuclear
envelope also defined by GO annotation, involving 231 genes. Figure 28b provides the
observed and permuted qq-plots associated with this list, and as expected, there was no
significant relationship between multiple gene products of the nuclear envelope and MI
(p=0.95).

Permutation analysis was also used to assess significance because the genes in the
lists are not summarized by a single SNP (e.g. with the minimum or median p-value), but
rather including the p-values from all SNPs genotyped in a given gene. While complicating
the qq-plots (e.g. Figure 27) by what can be substantial LD, keeping all typed SNPs in the
analysis ensures that susceptibility variants on different risk haplotypes are included.

CFTR, itself, is present on the apical membrane list and emphasizes the benefit of retaining
multiple associated SNPs. Figure 29a provides the gene-based perspective, where the red
line connects the p-values observed in CFTR and each black line represents the p-values
connected for one of 1000 simulated replicates in CFTR (simulated assuming no
association between MI and CFTR). The red line in Figure 5A is an extreme of the
distribution of black lines, despite only moderate individual p-values spanning CFTR
(Figure 29b). When each SNP-MI association was adjusted by CFTR mutation type (i.e.
using broad categories of mutations that lead to immature protein, truncated protein product
or aberrant transcripts) all association with MI in CFTR disappears i.e. the new observed
red line would sit at the bottom of the distribution of black lines, highlighting that there is
substantial allelic heterogeneity represented by SNPs tagging different CFTR mutations.
Thus different CFTR mutations provide variable susceptibility to MI, and this MI-CFTR
relationship would not be detected by single-SNP analysis on a genome-wide level.

The apical plasma membrane hypothesis was interrogated to determine whether it
increased power to detect individual genes that play a role in MI. To accomplish this goal
the SFDR (31) was uses as part of the GWAS-HD to reprioritize the genome according to
the hypothesis. Figures 25 and 30 illustrate that, in addition to SLC6A14, association
evidence with MI for gene SLC26A9 increased considerably using the GWAS-HD
approach, providing a q-value of 0.0007 when the expected false discovery rate is 0.07% among SNPs with q-value < 0.0007. Significance by GWAS-HD of 5 additional SNPs in SLC26A9 (Figure 25), some of which are not in LD (Figure 31) was also noted. When analysis is restricted to the subset of individuals who are homozygous for AF508, the most common CFTR mutation, multiple SNPs from SLC26A9 remained significant by GWAS-HD (q=0.0038) despite a smaller sample size, as does rs3788766 in SLC6A14 (q=0.0006). The permutation test of the apical membrane list also yielded significance (p=0.002). In addition, SLC9A3, ABCG8, and ATP2B2 from the AF508/AF508 sub-analysis each provide SFDR q-values < 0.05 (Figure 30).

In summary, the MI GWAS and GWAS-HD provides significant evidence that multiple genes present at the apical plasma membrane may contribute to the MI phenotype. As a result, multiple genes were prioritized for further study, many of which would have otherwise been designated as being of insufficient significance.

Modifiers for Multiple CF co-morbidities

A preliminary inspection of 998 Canadian patients on whom liver disease, diabetes, and MI data were known past the age range of typical onset, suggests a significant number of patients with both MI and CFRD (p=0.04), CFRD and CFLD (p<0.0001), but not MI and CFLD. In addition, evidence for common genetic modifiers between the various co-morbidities has been striking from initial GWAS results. For example, as in the MI GWAS one of the most strongly supported loci in the CFRD GWAS was SLC26A9 (p=5.7x10^{-6}), with consistency in both the alleles and direction of effect to the MI findings. Although SLC6A14, the strongest MI finding indicated only limited evidence of association with CFRD (p=0.08, rs3788766) and CFLD (p= 0.09), some overlap with lung disease findings existed. From the lung disease GWAS analysis, the third most significant locus included AGTR2, the gene neighbouring SLC6A14 on chromosome X. Further, it is notable that some of the SNPs with association evidence at this locus (e.g. rs6520219 with p<1x10^{-4}) may correspond to the promoter region of SLC6A14. Another interesting observation involves SLC9A3, which is also an apical plasma membrane constituent and is reported to be associated with lung disease in a CF candidate gene study. SNPs in SLC9A3 provided association evidence from both the lung disease GWAS and the MI GWAS (with p=0.0003
for lung and 0.0001 for MI for rs6864158 with the minor allele associated with a decreased risk of MI but with improved lung function, as expected.

**Determination of Modifiers for MI and lung disease**

A simple linear combination method that uses the average of two phenotype-specific association test statistics accounting for the baseline correlation between the two phenotypes being considered was used to identify loci that influence occurrence of MI and deteriorating lung function, CFRD and CFLD.

The statistic takes the form of \( \frac{\bar{T}_1 + T_2}{\sqrt{2(1 + p)}} \), where \( p \) is the correlation between the two phenotype-specific association test statistics. This statistic is normally distributed with mean 0 and variance 1 under the null of no association. A preliminary analysis was conducted applying this method to the combination of MI (yes/no) and deteriorating lung-function (quantitative trait as outlined in Taylor et al. appended) to look for common modifiers between the two co-morbidities. An empirical method was used to estimate the correlation value \( p \) between the two phenotypes by calculating the sample correlation between two vectors of \( T_i \) and \( T_2 \) after LD \( (r^2 < 0.2) \) and MAF \( (\text{MAF} > 0.05) \) pruning of the GWAS SNPs (94,737 SNPs left from the original 556,445 SNPs). This estimation method is justified by results of both our analytical derivation and simulation studies. The empirical correlation is approximately zero (-0.02), consistent with the observation that having MI does not lead to noticeably poorer lung function. Because it was anticipated that the effects of a SNP on the two phenotypes would be opposing, *i.e.* if the minor allele of a pleiotropic SNP increases occurrence of MI then it is reasonable that the same allele would correspond to poorer lung function, the above statistic was modified by changing the sign of one of the two phenotype-specific association statistics.

**Figure 33** shows the genome-wide qq-plot of the -log 10 (p-values) inferred from the combined statistic for detecting pleiotropic effects. **Figure 34** provides the top 16 SNPs (p-value \( \leq 10^{-5} \)) identified by this method. The qq-plot shows that a) the proposed method is accurate in that it does not increase the false positives and b) the tail distribution indicates multiple interesting results (**Figure 34**). The top ranked SNPs point us to a gene on chromosome 5, SLC9A3 a member of the solute carrier family present on the apical plasma membrane, and already identified in the MI GWAS-HD AF/AF sub-analysis.
Functional assessment of SLC6A14

SLC6A14 has been described as an electrogenic Na⁺ and Cl⁻ dependent amino acid transport system. Therefore, its function and its impact on overall trans-epithelial ion transport can be assessed by means of electrophysiological examination in an Ussing chamber.

Experiments were performed on primary human airway epithelial cells derived from lung explants of CF patients (F508del/F508del). In CF airway cells the dominant ion transport is luminal Na⁺-absorption via luminal Na⁺ channels. After blocking the Na⁺ channels with amiloride, the lumen-negative transepithelial potential difference (mV) decreased significantly. Further, in CF airway epithelial IBMX/Forskolin induced Cl⁻ secretion via CFTR Cl⁻ channels is reduced or absent. Under these conditions, apical application of the amino acid arginine induced an electrogenic dibasic amino acid transport which is characteristic for the SLC6A14 transport system (shown as lumen-negative change in mV). Addition of a specific CFTR inhibitor (CFTR172inh) decreased the lumen-negative potential difference. While addition of CFTR inhibitor did not other alternative Ca²⁺-dependent Cl⁻ channels (normal ATP response), this effect may reflect partial inhibition of the SLC6A14 system.

Modifiers of Meconium Ileus in Cystic Fibrosis

As described above, the North American CF Gene Modifier Consortium has accumulated a large patient collection, with 3,763 participants with 'severe' (pancreatic exocrine insufficient) CFTR genotypes and genome-wide genotype data of high quality at 543,927 SNPs. The definition of MI was consistent within the consortium and was recorded following rigorous chart review. The initial GWAS for MI used a generalized estimating equations (GEE) model to include collected sibling-pairs, and led to five genome-wide significant SNPs (P<5x10⁻⁸) from two regions that include SLC26A9 on chromosome 1 and SLC6A14 on chromosome X (Figure 35, Figure 36; sex-specific results in Figure 37). CFTR was not a significant confounder or effect modifier when incorporated in the GWAS (Figure 38 and Figure 39), indicating SLC6A14 and SLC26A9 are independent contributors to the MI phenotype.
The associations were successfully replicated at SLC6A14 (P=0.001) and SLC26A9 (P=0.0001) with MI in an independent combined sample from the North American collection and a French CF cohort (Figure 40).

The signal intensity plots of all of the associated SNPs reflected autosomal- and X-associated SNPs at SLC26A9 and SLC6A14, respectively. Imputation analysis using MACH identified the same regions of association as the genotyped SNPs (See Methods). The five associated SNPs for SLC6A14 and two for SLC26A9 are positioned just upstream of their respective transcription start sites such that binding of activating or repressing transcription factors may be affected (Figure 35b and 35c) as highlighted by the ENCODE data (data not shown). Neither SLC6A14 nor SLC26A9 coding regions exhibit significant evidence for CNVs; however, there is a noted gap in the genome sequence map > 10 kb upstream of the SLC26A9 locus.

The seven SNPs (Figure 36) of the two replicated genes account for < 5% of the MI variation, estimated by pseudo R-squared, likely reflecting substantial locus heterogeneity and low power to detect individual loci or SNPs given the available sample size. However, these genes were identified using the conventional GWAS designed for complex disease mapping, while in modifier gene studies critical disease information regarding the genetic etiology is often available and could be incorporated to identify modifier loci. In this modifier gene study of a recessive genetic disease, there is substantial information about the pathobiology of CF that can potentially lead to identification of a sizable collection of additional associated loci accounting for a much greater proportion of MI heritability. To do so, hypothesis-driven GWAS (GWAS-HD) was used to systematically prioritize the initial GWAS results. The GWAS-HD prioritization is based on the knowledge that a major source of CF pathophysiology is impaired fluid and electrolyte flux at the epithelial interfaces of many CF-affected organs including the airway, intestine, pancreas, liver and vas deferens. In these organs, the polarized epithelial layer forms a highly selective and tight barrier between organ and ductal interfaces. Transepithelial ‘function’ is achieved by cell polarization whereby many determinants and regulators of fluid, solute and ion transport reside at the apical membrane alongside CFTR, with contributing features from basolateral surfaces. In a mouse model it was shown that CFTR function in the gastrointestinal epithelium is critical for preventing intestinal obstructions. Thus, with loss
of CFTR, (genetic) variation in other apical membrane constituents could modify CF phenotypes, such as MI.

A list of 157 gene products (Figures 40 and 41) was annotated as localized to the apical plasma membrane using AmiGO with Gene Ontology data. CFTR and many solute transporters were included. However, the brush border membrane protein SLC6A14 was not listed, reflecting the high specialization of its corresponding intestinal cavity and a limitation of the GO annotation that we accepted without additional curation.

To test the apical hypothesis in the susceptibility to MI, GWAS-HD prioritized the genome by assigning SNPs of the apical genes to a high priority group versus all remaining SNPs of other genes. Two statistical procedures were implemented (Figure 42). First, using the prioritization, the stratified false discovery rate control (SFDR) was applied to the data to re-evaluate the initial association evidence for any given SNP at the genome-wide level. Then, a permutation-based test was used to determine the statistical significance of the apical hypothesis as a whole, testing all high priority SNPs jointly, to assess whether a preponderance of apical constituents contribute to MI susceptibility.

Even after the GWAS-HD analysis, SLC6A14 remained the gene with highest ranked SNPs for association with MI despite being assigned low priority (i.e. not an apical gene), reflecting the robustness of SFDR. In addition to SLC26A9, two genes, ATP2B2 and SLC9A3, showed association evidence with SNPs with q value< 0.05 (Figure 43), indicating that the extra power provided by GWAS-HD was needed to show statistical significance. A gene-based analysis (See Methods) also provided complementary association evidence for the four genes. After Bonferroni correction for the 156 tests performed (155 apical genes tagged by GWAS SNPs and SLC6A14), three genes remained significant, SLC6A14 (permutation P < 0.0001), SLC26A9 (P <0.0001) and SLC9A3 (P = 0.0001) (gene-based permutation P values in the French cohort are < 0.001, 0.235 and 0.017, respectively; Figure 41).

Next, testing the apical hypothesis as a whole (which excludes SLC6A14), GWAS-HD provided genome-wide significant evidence for association between MI and multiple constituents of the apical plasma membrane, permutation P = 0.0002, testing all 3,814 SNPs from 155 apical genes jointly and not subject to multiple hypothesis testing (Figures 40a and 40b). Even with the exclusion of SLC26A9 (and SLC6A14) to minimize any selection-bias concern, the apical hypothesis remained significant (P = 0.0058). Thus,
GWAS-HD established the involvement of other genes coding for apical constituents that we were not sufficiently powered to identify based on individual SNPs.

For comparison, a null hypothesis list of membrane-localized genes was constructed to test the GWAS-HD, for which we anticipated to see no relationship with MI. This list, also defined by GO annotation, comprised of all gene products present in the nuclear envelope (See Methods). The 224 nuclear envelope genes tagged by 3,537 GWAS SNPs showed no relationship with MI (permutation $P = 0.4639$; Figure 44).

The French cohort with genome-wide data provided independent validation of the apical hypothesis through genome-wide significant replication (permutation $P = 0.022$, testing all apical SNPs simultaneously, Figures 40b and 40c). There were a large number of genes with suggestive association evidence and their substantial consistency between the North American and French samples (Figure 41).

To determine the degree of involvement from the apical constituents and GWAS findings, Lasso was used to jointly analyze all 3,740 SNPs tagging the apical membrane genes (which include $SLC26A9$ and $SLC9A3$), and $SLC6A14$ (See Methods). Forty-eight SNPs spanning 36 different genes were retained by Lasso in the multivariate regression model (Figure 41). These SNPs jointly explain an appreciable amount (~17%, ~2.5-fold increase from the GWAS alone) of 'missing heritability' in MI, and indicated that the GWAS-HD framework yielded considerable additional information leading to interpretations beyond single SNP or single gene associations.

Methods

**Human Subjects.** Consent was obtained from all participants of the North American Cystic Fibrosis Gene Modifier Consortium (NACFGMC) with procedural approval from the Institutional Review Boards of Johns Hopkins University (JHU), the University of North Carolina at Chapel Hill (UNC) and Case Western Reserve University (CWRU) and the Research Ethics Board of The Hospital for Sick Children. Consent was also obtained for participants from France with procedural approval (CPP n°2004/15) and information collection approval by CNIL (n°04.404).

**Recruitment and Inclusions.** Cystic fibrosis (CF) patients and CF-related phenotypes including lung function and meconium ileus (MI) were collected by the NACFGMC. The
Genetic Modifier Study (GMS) included two sets of samples, one ascertained on the phenotype of extremes of lung disease (GMS-lung) and the other on the presence of CF-related severe liver disease (GMS-liver). The MI GWAS was restricted to subjects with ‘severe’ (pancreatic exocrine insufficient) CFTR genotypes and of Caucasian background (see quality control below). Participants (the 1,140 CF patients not used in the initial GWAS) for the North American replication corresponded to the continuing collections at all sites (351 from JHU Twin and Sibling Study (TSS), 448 from UNC/CWRU and 341 from CGS) with known MI status based on previously described criteria and as rigorously defined in source documentation and/or evidence of an abdominal scar.

There are 49 CF centers in France, caring for an estimated 5,000 to 6,000 CF patients. In 2006, prospective enrollment of CF patients was initiated in 38 of 49 CF centers. In January 2011, phenotypic information was available for 2,898 patients. Those selected for genotyping comprising the French replication cohort included 1,362 CF patients who were enrolled before June 2010, all of whom are over 6 years of age with two severe CFTR mutations and both parents born in a European country.

Genotyping. NACFGMC GWAS subjects were genotyped simultaneously using the Infinium HD Illumina 610-Quad BeadChip platform at McGill University and the Genome Quebec Innovation Center.

North American replication cohort. DNA was extracted from whole blood or transformed lymphocytes quantified with fluorimetry. Genotyping was performed with allele-specific fluorescent probes in Taqman® SNP Genotyping Assays (Custom or On-Demand; Applied Biosystems) as recommended using a 96-well format.

French replication cohort. DNA was obtained from whole blood and hybridized to the Illumina CNV370-Duo BeadChip for the first 299 patients (included before June 2009) and the Illumina 660W-Quad BeadChip for the remaining patients at the Centre National de Genotypage (CNG), Evry, France.

Quality control. GWAS genotypes. Samples with genotype missing rate >10%, heterozygosity proportion <28%, sex incongruity, and patients of non-Caucasian ancestry as determined by the principle component (PC) analysis using EIGENSTRAT were excluded. Using IBD estimates from PLINK and PREST-plus, twelve cryptic full-sib pairs
were identified and adjusted for relationships. Further, only one randomly selected individual from each of the 10 cryptic MZ pairs was retained, and parents of two cryptic parent-offspring pairs were deleted. In total, 3,763 samples were used for the analysis. SNPs with genotype call missing data rate >10%, MAF <2% were excluded, and 543,927 SNPs remained in the analysis.

North American replication genotypes. End-point fluorescence was measured with the plate reader component of the 7900HT Real Time PCR System (Applied Biosystems) and aided by Taqman® Genotyper software for allele discrimination with call rates >95%. Two percent of samples were run in duplicate and 1% of the samples corresponded to individuals used in the initial GWAS to assure quality control and permit assessment across genotyping platforms, respectively, with concordances >99%.

French replication genotypes. Patients with genotyping success rate <95%, sex incongruity, and pair-wise IBD estimates >40% were excluded, yielding a final set of 1,300 patients among which 1,232 had phenotype information for MI. SNPs present only on the CNV370-Duo chip were excluded from the analysis. SNPs with chip-wise missing data rate >10%, MAF <6% were excluded from the analysis. Overall, 554,792 SNPs were kept for the analysis, of which 256,756 were typed on both chips and 298,036 only on the 660W-Quad chip.

GO Annotation of the apical membrane constituent and nuclear envelope lists. The AmiGO tool\textsuperscript{13} (version 1.7) based on the Gene Ontology data was used to generate the two lists. A list of 157 apical genes was generated (retrieved March 28, 2010; GO:00163245) by the cell location search phrase "apical plasma membrane" with restriction to *Homo sapiens* (*SLC6A14* not on the list). In total 3,814 GWAS SNPs are within ± 10 kb of the boundaries of 155 genes (NCBI36/hg18); two genes are not tagged by any of the genotyped SNPs after QC. A list of 231 nuclear genes was generated (retrieved April 17, 2010; GO:0005635) by the cell location search phrase "nuclear envelope" with restriction to *Homo sapiens*. This list consisted of all gene products associated with the nuclear membrane. In total 3,537 GWAS SNPs are within ±10 kb of the boundaries of 224 genes.

Imputation. Using MACH, genotype imputation was conducted in two regions (*SLC6A14* on chromosome X and *SLC26A9* on chromosome 1) for the 3,763 subjects. The reference
sample was the 90 CEU subjects extracted from the EUR continental group of the 1000 genomes August 2010 release provided in the four-site (Broad Institute, Michigan University, Boston College and NCBI) merged dataset. Imputation yielded genotype data for 250 chromosome X SNPs and 2,639 chromosome 1 SNPs, among which 175 and 183 SNPs with estimated imputation accuracy >0.3 (using MACH’s R-squared accuracy measure) were considered for the association analysis. In SLC26A9, the best imputed SNP was only marginally more significant than any genotyped (6.23 x 10⁻⁹ vs. 9.88 x 10⁻⁹), while in SLC6A14 the minimum P value was provided by rs3788766, one of the genotyped SNPs.

**Statistical methods.** *Association analysis.* Generalized estimating equations (GEE⁹) was used for GWAS with an exchangeable correlation structure to account for the full-sib relationship in the data (Geeglm function in R, version 2.9.2). Genotypes were coded additively for autosomal SNPs and chromosome X SNPs in females. In males, 0 and 2 were used for chromosome X SNPs. A site covariate with four levels (CGS, GMS-lung, GMS-liver, and TSS) was included. Logistic regression in a sample of 3,199 unrelated individuals with a site covariate and the first seven principal components was also conducted, and results are consistent with the analysis of the full 3,763 subjects. (Therefore the PCs were not included in the subsequent analysis and permutation tests.) The French GWAS used logistic regression with additive genotype coding (PLINK v1.07 for autosomal SNPs and R for X chromosome SNPs).

*GWAS-HD, SFDR and Permutation.* GWAS-HD was used to accomplish two tasks: (1) to establish significance of individual SNPs at the genome-wide level after weighting according to a particular hypothesis, and (2) to test the significance of the hypothesis itself, by assessing whether the group of SNPs defined by the hypothesis display significantly smaller P values than would be expected under the null of no association.

To carry out the first task, GWAS SNPs were assigned to a high priority group (SNPs from the genes on the apical gene list) or a low priority group (all other SNPs). Stratified FDR control (SFDR) was then applied and q values were calculated separately in each group. Statistical significance at a given SNP was concluded if its q value was less than 0.05; each SNP was re-ranked genome-wide according to its new q value (the original GWAS P values were used to guide order if q values of two SNPs were identical).
To carry out the second task, that is to determine the statistical significance of the apical hypothesis involving 3,814 SNPs simultaneously (or 3,420 SNPs in the French replication cohort), the MI phenotype was permuted (to preserve the LD pattern between SNPs) within each consortium site and independently 10,000 times (or 1,000 in the French cohort). For each permutation sample, corresponding association analysis was performed, and a sum of the Wald association statistics of the 3,814 (or 3,420) SNPs was obtained. The empirical $P$ value for the significance of the apical hypothesis was calculated as the number of permuted samples whose sum statistics were larger than that in the observed data, divided by 10,000 (or 1,000).

Gene-based analysis. The analysis is similar to the permutation test above, but the sum statistic was obtained across all SNPs within $\pm$ 10 kb of the boundaries of each gene. In total, 156 gene-based permutation tests were performed (155 apical genes and SLC6A14), and a conservative Bonferroni adjusted statistical significance level is $\text{Error! Objects cannot be created from editing field codes.}$

Lasso. To determine which SNPS/genes jointly contribute to MI susceptibility, a multivariate analysis using penalized logistic regression (Lasso) was performed on 3,199 unrelated individuals (574 MI cases) extracted from the original 3,763 MI GWAS sample. The 3,814 SNPs from the apical plasma membrane list together with 15 SNPs within SLC6A14 were considered in the joint analysis (all SNPs with MAF > 2%). After removing 93 SNPs in perfect LD with one another (SNPs with $r^2 = 1$), a total of 3,740 SNPs were included as predictors in the Lasso analysis. The multivariate model also included the site covariate. The glmnet package in R was used to implement the Lasso. The default option in glmnet to standardize all predictors was turned off. The optimal value of the tuning parameter $\lambda$ was chosen based on 10-fold cross-validation (CV) to maximize the deviance.

Because the CV procedure randomly partitions the original data into training and testing sets, the optimal value of $\lambda$ varies depending on how the data is split; we therefore repeated the 10-fold CV 50 times, and determined the optimal value of $\lambda$ by examining the distribution of 50 $\lambda$ values and choosing the mode.

Estimating the phenotypic variance. Pseudo R-squared was used as an estimate of the phenotypic variance explained by the SNPs of interest. Calculations used the Irm function in R by regressing MI on SNPs in Figure 36 (-5%) and all SNPs selected by Lasso (-17%).
Incorporation by Reference

All publications, patents, and patent applications mentioned herein are hereby incorporated by reference in their entirety as if each individual publication, patent or patent application was specifically and individually indicated to be incorporated by reference. In case of conflict, the present application, including any definitions herein, will control.

Also incorporated by reference in their entirety are any polynucleotide and polypeptide sequences which reference an accession number correlating to an entry in a public database, such as those maintained by The Institute for Genomic Research (TIGR) on the world wide web at tigr.org, the National Center for Biotechnology Information (NCBI) on the world wide web at ncbi.nlm.nih.gov, or miRBase on the world wide web at microrna.sanger.ac.uk.

Equivalents

Those skilled in the art will recognize, or be able to ascertain using no more than routine experimentation, many equivalents to the specific embodiments of the invention described herein. Such equivalents are intended to be encompassed by the following claims.
What is claimed is:

1. A method of identifying a subject as having increased risk of severe lung disease comprising the step of detecting in a biological sample from the subject an allele of a single nucleotide polymorphism, wherein the allele of the single nucleotide polymorphism is selected from the group consisting of:
   a) a C allele at single nucleotide polymorphism rs12793173;
   b) an A allele at single nucleotide polymorphism rs403543;
   c) a C allele at single nucleotide polymorphism rs9268905;
   d) a G allele at single nucleotide polymorphism rs4760506;
   e) a T allele at single nucleotide polymorphism rs2883884;
   f) an A allele at single nucleotide polymorphism rs12188164;
   g) a C allele at single nucleotide polymorphism rs1645366; and
   h) any combination thereof,
   thereby identifying the subject as having increased risk of severe lung disease.

2. The method of claim 1, further comprising the step of determining whether the biological sample lacks a wild-type CFTR gene.

3. The method of claim 1, wherein the subject lacks a wild-type CFTR gene.

4. The method of claim 1, wherein the subject has or is suspected of having cystic fibrosis.

5. A method of identifying subject as a carrier of an allele of a single nucleotide polymorphism associated with severe lung disease comprising the step of detecting in a biological sample from the subject an allele of a single nucleotide polymorphism selected from the group consisting of:
   a) a C allele at single nucleotide polymorphism rs12793173;
   b) an A allele at single nucleotide polymorphism rs403543;
   c) a C allele at single nucleotide polymorphism rs9268905;
   d) a G allele at single nucleotide polymorphism rs4760506;
   e) a T allele at single nucleotide polymorphism rs2883884;
   f) an A allele at single nucleotide polymorphism rs12188164;
   g) a C allele at single nucleotide polymorphism rs1645366; and
   h) any combination thereof,
thereby identifying the subject as a carrier of an allele of a single nucleotide polymorphism associated with severe lung disease.

6. The method of claim 5, wherein the subject is or is suspected of being a carrier of a mutated CFTR gene.

5

7. A method of identifying a subject as having increased risk of severe lung disease comprising the step of detecting in a biological sample from the subject a variant in a gene selected from the group consisting of: EHF, APIP, MC3R, CASS4, AURKA, CBLN4, C20orf106 and CSTF1, thereby identifying the subject as having increased risk of severe lung disease.

10

8. The method of claim 7, further comprising the step of determining whether the biological sample lacks a wild-type CFTR gene.

9. The method of claim 7, wherein the subject lacks a wild-type CFTR gene.

10. The method of claim 7, wherein the subject has or is suspected of having cystic fibrosis.

15

11. A method of identifying subject as a carrier of an allele of a gene associated with severe lung disease comprising the step of detecting in a biological sample from the subject a variant in a gene selected from the group consisting of: EHF, APIP, MC3R, CASS4, AURKA, CBLN4, C20orf106 and CSTF1, thereby identifying the subject as a carrier of an allele of a single nucleotide polymorphism associated with severe lung disease.

20

12. The method of claim 11, further comprising the step of determining whether the biological sample comprises a CFTR gene mutation.

13. The method of claim 11, wherein the subject is or is suspected of being a carrier of a mutated CFTR gene.

14. The method of any one of claims 1-13, wherein the subject has at least one family member that has or is suspected of having cystic fibrosis.

15. The method of any one of claims 1-13, wherein the step of detecting comprises performing a hybridization assay.

16. The method of any one of claims 1-13, wherein the step of detecting comprises performing a nucleic acid amplification assay.

30

17. The method of any one of claims 1-13, wherein the step of detecting comprises performing a nucleic acid sequencing assay.
18. The method of any one of claims 1-13, further comprising the step of obtaining the biological sample from the subject.

19. The method of any one of claims 1-13, wherein the sample is selected from the group consisting of a tissue sample, a blood sample, a semen sample and a germ cell sample.

20. The method of any one of claims 1-13, wherein the subject is a human adult, a human child, a human fetus, a human embryo or a human fertilized cell.

21. A method of determining whether a test compound is a candidate therapeutic agent for reducing lung disease severity comprising:

   a) contacting a cell with the test compound; and

   b) detecting the expression by the cell of a gene product of a gene selected from the group consisting of: EHF, APIP, MC3R, CASS4, AURKA, CBLN4, C20orf106 and CSTF1;

   wherein a test compound that modulates the expression of the gene product is a candidate therapeutic agent for reducing lung disease severity.

22. The method of claim 21, wherein the gene product is an mRNA.

23. The method of claim 21, wherein the gene product is a protein.

24. The method of claim 23, wherein the protein is linked to a detectable moiety.

25. The method of claim 24, wherein the expression of the protein is detected by detecting the detectable moiety.

26. The method of any one of claims 21-25, wherein the agent is a small molecule, a polypeptide, an antibody or an inhibitory RNA molecule.

27. A method of reducing lung disease severity in a subject comprising administering to the subject a therapeutic agent that modulates the expression or activity of a gene product encoded by a gene selected from the group consisting of: EHF, APIP, MC3R, CASS4, AURKA, CBLN4, C20orf106 and CSTF1.

28. The method of claim 27, wherein the subject has or is suspected of having cystic fibrosis.

29. The method of claim 27 or 28, wherein the agent is a small molecule, a polypeptide, an antibody or an inhibitory RNA molecule.

30. The method of claim 27 or 28, wherein the agent reduces the expression or activity of the gene product.
31. A method of identifying a subject as having increased risk of having cystic fibrosis liver disease (CFLD) comprising the step of detecting in a biological sample from the subject an allele of a single nucleotide polymorphism, wherein the allele of the single nucleotide polymorphism is selected from the group consisting of:

a) a C allele at single nucleotide polymorphism rs914232;
b) a T allele at single nucleotide polymorphism rs2330183;
c) a T allele at single nucleotide polymorphism rs2838956;
d) a G allele at single nucleotide polymorphism rs1051266;
e) a T allele at single nucleotide polymorphism rs4819130;
f) a G allele at single nucleotide polymorphism rs3788190;
g) a T allele at single nucleotide polymorphism rs2236483;
h) a C allele at single nucleotide polymorphism rs2838950;
i) a G allele at single nucleotide polymorphism rs12483377;
j) a C allele at single nucleotide polymorphism rs3753019; and

k) any combination thereof,

thereby identifying the subject as having increased risk of having CFLD.

32. The method of claim 31, further comprising the step of determining whether the biological sample lacks a wild-type CFTR gene.

33. The method of claim 31, wherein the subject lacks a wild-type CFTR gene.

34. The method of claim 31, wherein the subject has or is suspected of having cystic fibrosis.

35. A method of identifying subject as a carrier of an allele of a single nucleotide polymorphism associated with CFLD comprising the step of detecting in a biological sample from the subject an allele of a single nucleotide polymorphism selected from the group consisting of:

a) a C allele at single nucleotide polymorphism rs914232;
b) a T allele at single nucleotide polymorphism rs2330183;
c) a T allele at single nucleotide polymorphism rs2838956;
d) a G allele at single nucleotide polymorphism rs1051266;
e) a T allele at single nucleotide polymorphism rs4819130;
f) a G allele at single nucleotide polymorphism rs3788190;
g) a T allele at single nucleotide polymorphism rs2236483;
h) a C allele at single nucleotide polymorphism rs2838950;
i) a G allele at single nucleotide polymorphism rs12483377;
j) a C allele at single nucleotide polymorphism rs3753019; and
k) any combination thereof,

thereby identifying the subject as a carrier of an allele of a single nucleotide polymorphism associated with CFLD.

36. The method of claim 35, wherein the subject is or is suspected of being a carrier of a mutated CFTR gene.
37. A method of identifying a subject as having increased risk of CFLD comprising the step of detecting in a biological sample from the subject a gene variant in SLC19A1 or COL18A1, thereby identifying the subject as having increased risk of CFLD.
38. The method of claim 37, further comprising the step of determining whether the biological sample lacks a wild-type CFTR gene.
39. The method of claim 37, wherein the subject lacks a wild-type CFTR gene.
40. The method of claim 37, wherein the subject has or is suspected of having cystic fibrosis.
41. A method of identifying subject as a carrier of an allele of a gene associated with CFLD comprising the step of detecting in a biological sample from the subject a gene variant in SLC19A1 or COL18A1, thereby identifying the subject as a carrier of an allele of a single nucleotide polymorphism associated with CFLD.
42. The method of claim 41, further comprising the step of determining whether the biological sample comprises a CFTR gene mutation.
43. The method of claim 41, wherein the subject is or is suspected of being a carrier of a mutated CFTR gene.
44. The method of any one of claims 31-43, wherein the subject has at least one family member that has or is suspected of having cystic fibrosis.
45. The method of any one of claims 31-43, wherein the step of detecting comprises performing a hybridization assay.
46. The method of any one of claims 31-43, wherein the step of detecting comprises performing a nucleic acid amplification assay.
47. The method of any one of claims 31-43, wherein the step of detecting comprises performing a nucleic acid sequencing assay.
48. The method of any one of claims 31-43, further comprising the step of obtaining the biological sample from the subject.
49. The method of any one of claims 31-43, wherein the sample is selected from the group consisting of a tissue sample, a blood sample, a semen sample and a germ cell sample.
50. The method of any one of claims 31-43, wherein the subject is a human adult, a human child, a human fetus, a human embryo or a human fertilized cell.
51. A method of determining whether a test compound is a candidate therapeutic agent for treating CFLD comprising:
   a) contacting a cell with the test compound; and
   b) detecting the expression by the cell of a gene product of SLC19A1 or COL18A1; wherein a test compound that modulates the expression of the gene product is a candidate therapeutic agent for treating CFLD.
52. The method of claim 51, wherein the gene product is an mRNA.
53. The method of claim 51, wherein the gene product is a protein.
54. The method of claim 53, wherein the protein is linked to a detectable moiety.
55. The method of claim 54, wherein the expression of the protein is detected by detecting the detectable moiety.
56. The method of any one of claims 51-55, wherein the agent is a small molecule, a polypeptide, an antibody or an inhibitory RNA molecule.
57. A method of treating CFLD in a subject comprising administering to the subject a therapeutic agent that modulates the expression or activity of a gene product of SLC19A1 or COL18A1.
58. The method of claim 57, wherein the subject has or is suspected of having cystic fibrosis.
59. The method of claim 57 or 58, wherein the agent is a small molecule, a polypeptide, an antibody or an inhibitory RNA molecule.
60. The method of claim 57 or 58, wherein the agent reduces the expression or activity of the gene product.
61. A method of identifying a subject as having increased risk of meconium ileus (MI) comprising the step of detecting in a biological sample from the subject an allele of a single
nucleotide polymorphism, wherein the allele of the single nucleotide polymorphism is selected from the group consisting of:

a) a C allele at single nucleotide polymorphism rs75 12462;
b) a G allele at single nucleotide polymorphism rs74 15921;
c) a G allele at single nucleotide polymorphism rs4077468;
d) a T allele at single nucleotide polymorphism rs4077469;
e) a G allele at single nucleotide polymorphism rs4077469;
f) an A allele at single nucleotide polymorphism rs7419153;
g) a T allele at single nucleotide polymorphism rs0179921;
h) a T allele at single nucleotide polymorphism rs4684689;
i) an A allele at single nucleotide polymorphism rs7563161;
j) a T allele at single nucleotide polymorphism rs3788766;
k) a C allele at single nucleotide polymorphism rs5905283;
l) a G allele at single nucleotide polymorphism rs0179921;
and

thereby identifying the subject as having increased risk of having MI.

62. The method of claim 61, further comprising the step of determining whether the biological sample lacks a wild-type CFTR gene.

63. The method of claim 61, wherein the subject lacks a wild-type CFTR gene.

64. The method of claim 61, wherein the subject has or is suspected of having cystic fibrosis.

65. A method of identifying subject as a carrier of an allele of a single nucleotide polymorphism associated with an increased risk of MI comprising the step of detecting in a biological sample from the subject an allele of a single nucleotide polymorphism selected from the group consisting of:

a) a C allele at single nucleotide polymorphism rs75 12462;
b) a G allele at single nucleotide polymorphism rs74 15921;
c) a G allele at single nucleotide polymorphism rs4077468;
d) a T allele at single nucleotide polymorphism rs4077469;
e) a G allele at single nucleotide polymorphism rs4077469;
f) an A allele at single nucleotide polymorphism rs7419153;
g) a T allele at single nucleotide polymorphism rs0179921;
h) a T allele at single nucleotide polymorphism rs4684689;
  i) an A allele at single nucleotide polymorphism rs17563161;
  j) a T allele at single nucleotide polymorphism rs3788766;
  k) a C allele at single nucleotide polymorphism rs5905283;
  l) a G allele at single nucleotide polymorphism rs12839137; and
  k) any combination thereof,

thereby identifying the subject as a carrier of an allele of a single nucleotide polymorphism associated with an increased risk of MI.

66. The method of claim 65, wherein the subject is or is suspected of being a carrier of a mutated CFTR gene.

67. A method of identifying a subject as having increased risk of MI comprising the step of detecting in a biological sample from the subject a variant in a gene selected from the group consisting of: SLC26A9, SLC6A14, SLC9A3, ABCG8 and ATP2B2, thereby identifying the subject as having increased risk of MI.

68. The method of claim 67, further comprising the step of determining whether the biological sample lacks a wild-type CFTR gene.

69. The method of claim 67, wherein the subject lacks a wild-type CFTR gene.

70. The method of claim 67, wherein the subject has or is suspected of having cystic fibrosis.

71. A method of identifying subject as a carrier of an allele of a gene associated with an increased risk of MI comprising the step of detecting in a biological sample from the subject a variant in a gene selected from the group consisting of: SLC26A9, SLC6A14, SLC9A3, ABCG8 and ATP2B2, thereby identifying the subject as a carrier of an allele of a single nucleotide polymorphism associated with an increased risk of MI.

72. The method of claim 71, further comprising the step of determining whether the biological sample comprises a CFTR gene mutation.

73. The method of claim 71, wherein the subject is or is suspected of being a carrier of a mutated CFTR gene.

74. The method of any one of claims 61-73, wherein the subject has at least one family member that has or is suspected of having cystic fibrosis.

75. The method of any one of claims 61-73, wherein the step of detecting comprises performing a hybridization assay.
76. The method of any one of claims 61-73, wherein the step of detecting comprises performing a nucleic acid amplification assay.

77. The method of any one of claims 61-73, wherein the step of detecting comprises performing a nucleic acid sequencing assay.

78. The method of any one of claims 61-73, further comprising the step of obtaining the biological sample from the subject.

79. The method of any one of claims 61-73, wherein the sample is selected from the group consisting of a tissue sample, a blood sample, a semen sample and a germ cell sample.

80. The method of any one of claims 61-73, wherein the subject is a human adult, a human child, a human fetus, a human embryo or a human fertilized cell.

81. A method of determining whether a test compound is a candidate therapeutic agent for treating MI comprising:
   a) contacting a cell with the test compound; and
   b) detecting the expression by the cell of a gene product of a gene selected from the group consisting of: SLC26A9, SLC6A14, SLC9A3, ABCG8 and ATP2B2;
   wherein a test compound that modulates the expression of the gene product is a candidate therapeutic agent for treating MI.

82. The method of claim 81, wherein the gene product is an mRNA.

83. The method of claim 81, wherein the gene product is a protein.

84. The method of claim 83, wherein the protein is linked to a detectable moiety.

85. The method of claim 84, wherein the expression of the protein is detected by detecting the detectable moiety.

86. The method of any one of claims 81-85, wherein the agent is a small molecule, a polypeptide, an antibody or an inhibitory RNA molecule.

87. A method of treating MI in a subject comprising administering to the subject a therapeutic agent that modulates the expression or activity of a gene product of a gene selected from the group consisting of: SLC26A9, SLC6A14, SLC9A3, ABCG8 and ATP2B2.

88. The method of claim 87, wherein the subject has or is suspected of having cystic fibrosis.
89. The method of claim 87 or 88, wherein the agent is a small molecule, a polypeptide, an antibody or an inhibitory RNA molecule.

90. The method of claim 87 or 88, wherein the agent reduces the expression or activity of the gene product.
## Figure 1

<table>
<thead>
<tr>
<th>SNP</th>
<th>Forward Strand Flanking Sequence</th>
<th>Chromosome</th>
<th>GI Number</th>
<th>Position *</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs12793173</td>
<td>TGGTTGTTTCTATTAGAAGACGAGAGCAGG</td>
<td>11</td>
<td>224514737</td>
<td>34774204</td>
</tr>
<tr>
<td>rs1403543</td>
<td>TTAACACTGATTGTGGCAAAACTCTCTCT</td>
<td>X</td>
<td>224589822</td>
<td>115302192</td>
</tr>
<tr>
<td>rs9268905</td>
<td>GACCTGTTGAGGAGCCAGAGAGAGGAG</td>
<td>6</td>
<td>224589818</td>
<td>32432077</td>
</tr>
<tr>
<td>rs4760506</td>
<td>GAGAGTGGATGAGCAGAGGAGAGAGAG</td>
<td>12</td>
<td>224589803</td>
<td>9333050</td>
</tr>
<tr>
<td>rs12838848</td>
<td>GACGACCACTCAGCAGGAGAGAGAGGA</td>
<td>14</td>
<td>224589805</td>
<td>70517183</td>
</tr>
<tr>
<td>rs12188161</td>
<td>GACTTCTCCTGTTCTGCTGTTCTGCTGCT</td>
<td>5</td>
<td>224589817</td>
<td>428236</td>
</tr>
<tr>
<td>rs11645366</td>
<td>CTAACCTGTGAGGAGGAGGAGAGAGAG</td>
<td>16</td>
<td>224589807</td>
<td>62377153</td>
</tr>
<tr>
<td>rs914232</td>
<td>CATACACATTGAGGGAGAGAGAGAGAGA</td>
<td>21</td>
<td>224589813</td>
<td>46952750</td>
</tr>
<tr>
<td>rs2330183</td>
<td>TAGAGGGCTGAGGAGGAGAGAGAGAGAG</td>
<td>21</td>
<td>224589813</td>
<td>46953292</td>
</tr>
<tr>
<td>rs2838956</td>
<td>GTGAGGGCTGAGGAGGAGAGAGAGAGAG</td>
<td>21</td>
<td>224589813</td>
<td>46954024</td>
</tr>
<tr>
<td>rs1051266</td>
<td>TAGAGGGCTGAGGAGGAGAGAGAGAGAG</td>
<td>21</td>
<td>224589813</td>
<td>46957794</td>
</tr>
<tr>
<td>rs4819130</td>
<td>TAGAGGGCTGAGGAGGAGAGAGAGAGAG</td>
<td>21</td>
<td>224589813</td>
<td>46958299</td>
</tr>
<tr>
<td>rs3788190</td>
<td>TAGAGGGCTGAGGAGGAGAGAGAGAGAG</td>
<td>21</td>
<td>224589813</td>
<td>46960398</td>
</tr>
<tr>
<td>rs2236483</td>
<td>TAGAGGGCTGAGGAGGAGAGAGAGAGAG</td>
<td>21</td>
<td>224589813</td>
<td>46962054</td>
</tr>
<tr>
<td>rs2838950</td>
<td>TAGAGGGCTGAGGAGGAGAGAGAGAGAG</td>
<td>21</td>
<td>224589813</td>
<td>46962697</td>
</tr>
<tr>
<td>rs12483377</td>
<td>TAGAGGGCTGAGGAGGAGAGAGAGAGAG</td>
<td>21</td>
<td>224589813</td>
<td>46931109</td>
</tr>
<tr>
<td>rs3753019</td>
<td>TAGAGGGCTGAGGAGGAGAGAGAGAGAG</td>
<td>21</td>
<td>224589813</td>
<td>46924785</td>
</tr>
<tr>
<td>rs7512462</td>
<td>TAGAGGGCTGAGGAGGAGAGAGAGAGAG</td>
<td>21</td>
<td>224589800</td>
<td>205890959</td>
</tr>
<tr>
<td>rs7415921</td>
<td>TAGAGGGCTGAGGAGGAGAGAGAGAGAG</td>
<td>21</td>
<td>224589800</td>
<td>205910883</td>
</tr>
<tr>
<td>rs4077468</td>
<td>TAGAGGGCTGAGGAGGAGAGAGAGAGAG</td>
<td>21</td>
<td>224589800</td>
<td>205914757</td>
</tr>
<tr>
<td>rs4077469</td>
<td>TAGAGGGCTGAGGAGGAGAGAGAGAGAG</td>
<td>21</td>
<td>224589800</td>
<td>205914885</td>
</tr>
<tr>
<td>rs12047830</td>
<td>TAGAGGGCTGAGGAGGAGAGAGAGAGAG</td>
<td>21</td>
<td>224589800</td>
<td>205916699</td>
</tr>
<tr>
<td>rs7419153</td>
<td>TAGAGGGCTGAGGAGGAGAGAGAGAGAG</td>
<td>21</td>
<td>224589800</td>
<td>205917309</td>
</tr>
<tr>
<td>rs10179921</td>
<td>TAGAGGGCTGAGGAGGAGAGAGAGAGAG</td>
<td>21</td>
<td>224589811</td>
<td>44068291</td>
</tr>
<tr>
<td>rs4684689</td>
<td>TAGAGGGCTGAGGAGGAGAGAGAGAGAG</td>
<td>3</td>
<td>224589815</td>
<td>10448426</td>
</tr>
<tr>
<td>rs17563161</td>
<td>TAGAGGGCTGAGGAGGAGAGAGAGAGAG</td>
<td>5</td>
<td>224589817</td>
<td>497624</td>
</tr>
<tr>
<td>rs3787666</td>
<td>TAGAGGGCTGAGGAGGAGAGAGAGAGAG</td>
<td>X</td>
<td>224589822</td>
<td>115566839</td>
</tr>
<tr>
<td>rs5905283</td>
<td>TAGAGGGCTGAGGAGGAGAGAGAGAGAG</td>
<td>X</td>
<td>224589822</td>
<td>115565881</td>
</tr>
<tr>
<td>rs12839137</td>
<td>TAGAGGGCTGAGGAGGAGAGAGAGAGAG</td>
<td>X</td>
<td>224589822</td>
<td>115565550</td>
</tr>
</tbody>
</table>

a. NCBI build 37.2.
Figure 2

Characteristics of patients enrolled by the three studies comprising the North American CF Gene Modifier Consortium

<table>
<thead>
<tr>
<th></th>
<th>Genetic Modifier Study (GMS)</th>
<th>Canadian Consortium for Genetic Studies (CGS)</th>
<th>Twins &amp; Sibs Study (TSS)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lead Institution(s)</strong></td>
<td>Univ. of North Carolina/Case Western</td>
<td>Univ. of Toronto</td>
<td>Johns Hopkins</td>
</tr>
<tr>
<td><strong>Design</strong></td>
<td>Extremes-of-Phenotype Unrelated</td>
<td>Population-Based Unrelated</td>
<td>Family-Based</td>
</tr>
<tr>
<td><strong>Type of Evidence</strong></td>
<td>Association</td>
<td>Association</td>
<td>Linkage and association</td>
</tr>
<tr>
<td><strong>Number of patients</strong></td>
<td>1,137</td>
<td>1,357</td>
<td>973* (486 sibling pairs)</td>
</tr>
<tr>
<td></td>
<td>Severe (n = 406)</td>
<td>Mild (n = 731)</td>
<td></td>
</tr>
<tr>
<td><strong>Age</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean ± SD (yrs)</td>
<td>15.2 ± 4.6</td>
<td>27.5 ± 9.8</td>
<td>18.5 ± 9.5</td>
</tr>
<tr>
<td>Range (yrs)</td>
<td>8-25</td>
<td>15-56</td>
<td>6-49</td>
</tr>
<tr>
<td>Male n (%)</td>
<td>194 (47.8%)</td>
<td>405 (55.4%)</td>
<td>734 (54.1%)</td>
</tr>
<tr>
<td>Caucasian n (%)</td>
<td>1,137 (100.0%)</td>
<td>1,180 (87.0%)</td>
<td>898 (92.3%)</td>
</tr>
<tr>
<td>$F508del/F508del$ n (%)</td>
<td>1,137 (100.0%)</td>
<td>841 (62.0%)</td>
<td>557 (57.3%)</td>
</tr>
<tr>
<td>Pancreatic Exocrine Insufficient n (%)</td>
<td>1,137 (100.0%)</td>
<td>1,357 (100.0%)</td>
<td>973 (100.0%)</td>
</tr>
</tbody>
</table>

*420 two-sib families, 20 three-sib families, 1 four-sib family and 69 singletons.

*Based on self-identified ancestry and principal components analysis.
Figure 3

(a) Histogram showing the distribution of Consortium Quantitative Phenotype for different groups:
- GMS "Severe" (n=406)
- GMS "Mild" (n=731)
- CGS (n=1,357)

(b) Histogram showing the distribution of Consortium Quantitative Phenotype for TSS (n=973)
Table 1: Significant and suggestive association results for GMS and CGS, with replication values for TSS.

<table>
<thead>
<tr>
<th>SNP</th>
<th>Chr</th>
<th>Base pair</th>
<th>Nearest gene</th>
<th>Category</th>
<th>Risk allele</th>
<th>Non-risk allele</th>
<th>(Minor allele) Freq.</th>
<th>GMS F508del/F508del coef.</th>
<th>CGS F508del coef.</th>
<th>CGS All coef.</th>
<th>Analysis with max significance</th>
<th>GMS+CGS F508del</th>
<th>P-value: GMS+CGS F508del</th>
<th>P-value: GMS+CGS F508del</th>
<th>P-value: GMS+CGS All</th>
<th>P-value: GMS+CGS All</th>
<th>P-value: GMS+CGS TSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs12793173</td>
<td>11</td>
<td>34,790,780</td>
<td>APIP/EHF</td>
<td>Significant</td>
<td>C</td>
<td>T</td>
<td>(C) 0.24</td>
<td>0.16</td>
<td>0.20</td>
<td>0.12</td>
<td>GMS+CGS F508del/F508del</td>
<td>3.34E-08</td>
<td>1.76E-06</td>
<td>0.906</td>
<td>3.34E-08</td>
<td>1.76E-06</td>
<td>0.006</td>
</tr>
<tr>
<td>rs1403543</td>
<td>23</td>
<td>115,216,220</td>
<td>AGTR2</td>
<td>Suggestive</td>
<td>A</td>
<td>G</td>
<td>(G) 0.49</td>
<td>0.22</td>
<td>0.07</td>
<td>0.11</td>
<td>GMS+CGS All</td>
<td>1.61E-05</td>
<td>2.58E-07</td>
<td>0.053</td>
<td>1.61E-05</td>
<td>2.58E-07</td>
<td>0.053</td>
</tr>
<tr>
<td>rs9268905</td>
<td>6</td>
<td>32,540,055</td>
<td>HLA-DRA</td>
<td>Suggestive</td>
<td>C</td>
<td>G</td>
<td>(C) 0.32</td>
<td>0.16</td>
<td>0.10</td>
<td>0.12</td>
<td>GMS+CGS All</td>
<td>1.42E-05</td>
<td>2.81E-07</td>
<td>0.032</td>
<td>1.42E-05</td>
<td>2.81E-07</td>
<td>0.032</td>
</tr>
<tr>
<td>rs4760506</td>
<td>12</td>
<td>91,857,181</td>
<td>EE1A</td>
<td>Suggestive</td>
<td>G</td>
<td>A</td>
<td>(A) 0.46</td>
<td>0.16</td>
<td>0.10</td>
<td>0.10</td>
<td>GMS+CGS All</td>
<td>6.77E-06</td>
<td>8.56E-07</td>
<td>0.594</td>
<td>6.77E-06</td>
<td>8.56E-07</td>
<td>0.594</td>
</tr>
<tr>
<td>rs12883884</td>
<td>14</td>
<td>69,586,936</td>
<td>SLC8A3</td>
<td>Suggestive</td>
<td>T</td>
<td>G</td>
<td>(G) 0.39</td>
<td>0.12</td>
<td>0.15</td>
<td>0.12</td>
<td>GMS+CGS All</td>
<td>1.20E-06</td>
<td>9.56E-07</td>
<td>0.223</td>
<td>1.20E-06</td>
<td>9.56E-07</td>
<td>0.223</td>
</tr>
<tr>
<td>rs12188164</td>
<td>5</td>
<td>481,236</td>
<td>AHRR</td>
<td>Suggestive</td>
<td>A</td>
<td>C</td>
<td>(A) 0.38</td>
<td>0.08</td>
<td>0.12</td>
<td>0.15</td>
<td>GMS+CGS All</td>
<td>5.92E-04</td>
<td>1.34E-06</td>
<td>0.136</td>
<td>5.92E-04</td>
<td>1.34E-06</td>
<td>0.136</td>
</tr>
<tr>
<td>rs11645366</td>
<td>16</td>
<td>60,934,654</td>
<td>CDH8</td>
<td>Suggestive</td>
<td>C</td>
<td>T</td>
<td>(T) 0.23</td>
<td>0.17</td>
<td>0.13</td>
<td>0.13</td>
<td>GMS+CGS All</td>
<td>1.23E-05</td>
<td>1.52E-06</td>
<td>0.182</td>
<td>1.23E-05</td>
<td>1.52E-06</td>
<td>0.182</td>
</tr>
</tbody>
</table>

1NCBI build 36.
2Significant and suggestive imply $P < (0.05/570725) = 8.76 \times 10^{-5}$ or $P < (1/570725) = 1.75 \times 10^{-6}$, respectively, for at least one analysis (GMS+CGS F508del/F508del or GMS+CGS All).
3Alleles indexed to the forward strand of NCBI build 36; the risk allele is the allele associated with worse lung function.
4Minor allele frequencies are listed for all GMS+CGS All. Study-specific MAFs are provided in Supplementary Table 1.
5Coefficients refer to the average reduction in the Consortium lung phenotype for each copy of the risk allele.
6TSS direction-consistent association p-value, for TSS F508del/F508del only, or TSS All Patients, selected according to the GMS+CGS result with maximum significance.
7F508del/F508del
### GMS

<table>
<thead>
<tr>
<th>SNP</th>
<th>Chr</th>
<th>Base Pair</th>
<th>Nearest Gene</th>
<th>Combined Evidence</th>
<th>GMS Patient Set</th>
<th>HWE</th>
<th>original SNP coef</th>
<th>Coefficients (P-values) after additional covariate adjustment</th>
<th>SNPl</th>
<th>Sex</th>
<th>BMZ Score</th>
<th>TGFl</th>
<th>IFRDl</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs12793173</td>
<td>11</td>
<td>34,790,789</td>
<td>SHFFAPR</td>
<td>Significant</td>
<td>F508del/F508del</td>
<td>0.57</td>
<td>0.16</td>
<td>0.15 (2.66E-04)</td>
<td>0.34 (1.46E-04)</td>
<td>0.43 (4.66E-04)</td>
<td>0.12 (7.43E-04)</td>
<td>0.04 (3.78E-04)</td>
<td></td>
</tr>
<tr>
<td>rs1403542</td>
<td>26</td>
<td>127,216,220</td>
<td>AGTR2</td>
<td>Suggestive</td>
<td>F506del/F506del</td>
<td>0.39</td>
<td>0.22</td>
<td>0.18 (7.37E-06)</td>
<td>0.24 (6.44E-06)</td>
<td>0.43 (1.95E-06)</td>
<td>0.12 (9.11E-06)</td>
<td>0.04 (4.43E-06)</td>
<td></td>
</tr>
<tr>
<td>rs22840656</td>
<td>6</td>
<td>32,646,026</td>
<td>HLA-DRB1</td>
<td>Suggestive</td>
<td>F506del/F506del</td>
<td>0.38</td>
<td>0.20</td>
<td>0.14 (7.33E-05)</td>
<td>0.30 (3.67E-05)</td>
<td>0.43 (1.95E-05)</td>
<td>0.12 (9.11E-05)</td>
<td>0.05 (4.43E-05)</td>
<td></td>
</tr>
<tr>
<td>rs74650506</td>
<td>12</td>
<td>9,652,121</td>
<td>EEA1</td>
<td>Suggestive</td>
<td>F506del/F506del</td>
<td>0.11</td>
<td>0.16</td>
<td>0.11 (4.53E-04)</td>
<td>0.16 (4.84E-04)</td>
<td>0.43 (1.52E-04)</td>
<td>0.12 (8.16E-04)</td>
<td>0.04 (4.47E-04)</td>
<td></td>
</tr>
<tr>
<td>rs128983884</td>
<td>14</td>
<td>69,586,936</td>
<td>SLC2A3</td>
<td>Suggestive</td>
<td>F506del/F506del</td>
<td>0.18</td>
<td>0.12</td>
<td>0.10 (5.79E-03)</td>
<td>0.16 (1.33E-03)</td>
<td>0.43 (2.75E-03)</td>
<td>0.12 (6.45E-03)</td>
<td>0.04 (1.68E-03)</td>
<td></td>
</tr>
<tr>
<td>rs12181164</td>
<td>5</td>
<td>481,236</td>
<td>AHRR</td>
<td>Suggestive</td>
<td>F506del/F506del</td>
<td>0.94</td>
<td>0.08</td>
<td>0.07 (2.77E-03)</td>
<td>0.17 (6.89E-04)</td>
<td>0.43 (1.52E-04)</td>
<td>0.12 (8.16E-04)</td>
<td>0.05 (4.47E-04)</td>
<td></td>
</tr>
<tr>
<td>rs116453065</td>
<td>10</td>
<td>50,394,654</td>
<td>CDH18</td>
<td>Suggestive</td>
<td>F506del/F506del</td>
<td>0.16</td>
<td>0.17</td>
<td>0.12 (3.02E-03)</td>
<td>0.16 (1.39E-04)</td>
<td>0.43 (4.21E-04)</td>
<td>0.12 (1.17E-03)</td>
<td>0.04 (4.80E-04)</td>
<td></td>
</tr>
<tr>
<td>rs60266090</td>
<td>76</td>
<td>53,228,948</td>
<td>C3/Na/KC3R</td>
<td>Significant</td>
<td>F506del/F506del</td>
<td>0.85</td>
<td>0.14</td>
<td>0.11 (2.00E-02)</td>
<td>0.16 (3.13E-03)</td>
<td>0.43 (2.72E-03)</td>
<td>0.12 (8.14E-04)</td>
<td>0.04 (3.70E-04)</td>
<td></td>
</tr>
</tbody>
</table>

*SNP Identiﬁed by 3G.
**Coefficients refer to the average reduction in the Consortium lung phenotype for each copy of the SNP risk allele, adjusted for covariates sex + PCs (all coefﬁcients except for those corresponding to rs6024460 are repeated from Table 1).
**Coefficients refer to the average reduction in the Consortium lung phenotype for each copy of the SNP risk allele, adjusted for additional covariates BMZ, TGFl, and IFRDl.
**Coefficient is for female status, i.e. positive implies females have higher average phenotype.
*See Online Methods, under Additional Phenotyping.
**Coefficient is for the codon 10 “CC” genotype of TGFl.
**Coefficient is for the “CT” genotype of IFRDl.
*Genome-wide signiﬁcant in association scan of GMS+CGS F508del/F508del.
**Suggestive” implies expected less than once per genotype scan in the GMS+CGS All association scan.
*Genome-wide signiﬁcant linkage in region, followed by signiﬁcant regional association in GMS+CGS F508del/F508del.
Figure 9
Figure 14

IBD0  IBD1  IBD2

[Scatter plots showing data distribution for IBD0, IBD1, and IBD2]
<table>
<thead>
<tr>
<th>Chr</th>
<th>SNP</th>
<th>Base Pair</th>
<th>GMS+CGS</th>
<th>FDR q-value&lt;sup&gt;a&lt;/sup&gt;</th>
<th>FDR rank</th>
<th>WFDI q-value&lt;sup&gt;b&lt;/sup&gt;</th>
<th>WFDI rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>rs931386</td>
<td>9716289</td>
<td>5.08E-06</td>
<td>0.0124</td>
<td>7</td>
<td>0.0389</td>
<td>16</td>
</tr>
<tr>
<td>11</td>
<td>rs11032829</td>
<td>34705078</td>
<td>2.29E-06</td>
<td>0.008</td>
<td>3</td>
<td>0.0277</td>
<td>8</td>
</tr>
<tr>
<td>11</td>
<td>rs7924717</td>
<td>9473907</td>
<td>2.76E-06</td>
<td>0.008</td>
<td>1</td>
<td>0.0218</td>
<td>6</td>
</tr>
<tr>
<td>11</td>
<td>rs10466455</td>
<td>34737512</td>
<td>8.66E-07</td>
<td>0.0124</td>
<td>6</td>
<td>0.0375</td>
<td>14</td>
</tr>
<tr>
<td>11</td>
<td>rs7929679</td>
<td>34784425</td>
<td>1.47E-06</td>
<td>0.008</td>
<td>5</td>
<td>0.0282</td>
<td>10</td>
</tr>
<tr>
<td>11</td>
<td>rs10385112</td>
<td>34787019</td>
<td>1.56E-07</td>
<td>0.008</td>
<td>2</td>
<td>0.0277</td>
<td>7</td>
</tr>
<tr>
<td>11</td>
<td>rs525202</td>
<td>94778524</td>
<td>1.34E-07</td>
<td>0.008</td>
<td>4</td>
<td>0.0277</td>
<td>9</td>
</tr>
<tr>
<td>20</td>
<td>rs7265042</td>
<td>53790816</td>
<td>1.14E-03</td>
<td>0.7852</td>
<td>1976</td>
<td>0.0459</td>
<td>18</td>
</tr>
<tr>
<td>20</td>
<td>rs8098782</td>
<td>53791974</td>
<td>1.94E-08</td>
<td>0.7349</td>
<td>865</td>
<td>0.0459</td>
<td>17</td>
</tr>
<tr>
<td>20</td>
<td>rs910668</td>
<td>53794753</td>
<td>1.09E-03</td>
<td>0.5029</td>
<td>175</td>
<td>0.015</td>
<td>2</td>
</tr>
<tr>
<td>20</td>
<td>rs6092176</td>
<td>5379109</td>
<td>1.51E-03</td>
<td>0.5581</td>
<td>255</td>
<td>0.015</td>
<td>5</td>
</tr>
<tr>
<td>20</td>
<td>rs6092179</td>
<td>53812440</td>
<td>1.93E-04</td>
<td>0.484</td>
<td>154</td>
<td>0.015</td>
<td>1</td>
</tr>
<tr>
<td>20</td>
<td>rs6024457</td>
<td>53815962</td>
<td>1.61E-04</td>
<td>0.7553</td>
<td>1116</td>
<td>0.0353</td>
<td>13</td>
</tr>
<tr>
<td>20</td>
<td>rs8125625</td>
<td>53820352</td>
<td>2.49E-04</td>
<td>0.516</td>
<td>207</td>
<td>0.015</td>
<td>3</td>
</tr>
<tr>
<td>20</td>
<td>rs8024454</td>
<td>53826840</td>
<td>2.56E-04</td>
<td>0.7615</td>
<td>1340</td>
<td>0.0381</td>
<td>15</td>
</tr>
<tr>
<td>20</td>
<td>rs8024460</td>
<td>53828948</td>
<td>1.34E-04</td>
<td>0.7349</td>
<td>854</td>
<td>0.0296</td>
<td>12</td>
</tr>
<tr>
<td>20</td>
<td>rs11907114</td>
<td>53862354</td>
<td>2.79E-08</td>
<td>0.554</td>
<td>250</td>
<td>0.015</td>
<td>4</td>
</tr>
<tr>
<td>20</td>
<td>rs1325022</td>
<td>54277432</td>
<td>1.16E-03</td>
<td>0.7344</td>
<td>824</td>
<td>0.0296</td>
<td>11</td>
</tr>
</tbody>
</table>

<sup>a</sup> Benjamini-Hochberg approach based on association P-value.
<sup>b</sup> Weighted FDR using combined linkage information and association P-values.

Rows in bold indicate the top ranked SNPs before incorporating linkage evidence (rs7924717 on chromosome 11) and after (rs6092179 on chromosome 20).
Multiple Stimuli
Altered Immune State
Proliferative stimuli
Oxidative Stress
Inflammatory Factors

Stellate Cell

Migration and
Proliferation of HSC
HFG
IL6

Disruption of α₁-antitrypsin (Z allele)
Hepatocyte

Disruption of CFTR expression
Cholangiocyte

Abnormal Bile flow

Increases in
TGFB1, TGFB2,
SERPINE1 (PAI-1),
EDN1, TIMP1 procollagen

Net increases in the deposition of the extracellular matrix, leading to fibrosis & cirrhosis
<table>
<thead>
<tr>
<th>Consortium Center/Site</th>
<th>CFTR Status</th>
<th>MI Status</th>
<th>Gender (separately for cases and controls)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\Delta F508/\Delta F508$</td>
<td>Others</td>
<td>Cases (MI)</td>
</tr>
<tr>
<td>CGS (Canadian Gene Modifier Study Population Samples)</td>
<td>1593</td>
<td>967</td>
<td>626</td>
</tr>
<tr>
<td>GMS (North Carolina/Case Western Reserve Extreme Phenotype Study)</td>
<td>1193</td>
<td>1193</td>
<td>0</td>
</tr>
<tr>
<td>TSS (Johns Hopkins University Family-based Twin and Sibling Study)</td>
<td>869</td>
<td>515</td>
<td>354</td>
</tr>
<tr>
<td>Total</td>
<td>3655</td>
<td>2675</td>
<td>980</td>
</tr>
</tbody>
</table>
### Figure 23

<table>
<thead>
<tr>
<th>Sample</th>
<th>Sex</th>
<th>Genotype (# minor alleles, C)</th>
<th>Genotype Count</th>
<th>HWE p-value*</th>
<th>OR (95% CI)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>MI</td>
<td>No MI</td>
<td>MI</td>
<td>No MI</td>
<td>N/A</td>
</tr>
<tr>
<td>ALL</td>
<td>Female</td>
<td>0</td>
<td>146</td>
<td>510</td>
<td>0.25</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>114</td>
<td>666</td>
<td>0.59 (0.45, 0.78)</td>
<td>0.00019</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>34</td>
<td>210</td>
<td>0.52 (0.34, 0.80)</td>
<td>0.00266</td>
</tr>
<tr>
<td>Male</td>
<td></td>
<td>0</td>
<td>243</td>
<td>937</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>80</td>
<td>715</td>
<td>0.44 (0.33, 0.58)</td>
<td>3.3e-09</td>
</tr>
<tr>
<td>CGS only</td>
<td>Female</td>
<td>0</td>
<td>58</td>
<td>224</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>52</td>
<td>291</td>
<td>0.68 (0.45, 1.02)</td>
<td>0.064</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>17</td>
<td>87</td>
<td>0.72 (0.40, 1.29)</td>
<td>0.268</td>
</tr>
<tr>
<td>Male</td>
<td></td>
<td>0</td>
<td>99</td>
<td>409</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>34</td>
<td>322</td>
<td>0.44 (0.29, 0.66)</td>
<td>9.2e-05</td>
</tr>
<tr>
<td>TSS only</td>
<td>Female</td>
<td>0</td>
<td>45</td>
<td>111</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>25</td>
<td>166</td>
<td>0.37 (0.21, 0.64)</td>
<td>0.00047</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>10</td>
<td>52</td>
<td>0.41 (0.17, 1.00)</td>
<td>0.05071</td>
</tr>
<tr>
<td>Male</td>
<td></td>
<td>0</td>
<td>62</td>
<td>212</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>16</td>
<td>170</td>
<td>0.33 (0.18, 0.59)</td>
<td>0.00018</td>
</tr>
<tr>
<td>GMS only</td>
<td>Female</td>
<td>0</td>
<td>43</td>
<td>175</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>37</td>
<td>209</td>
<td>0.72 (0.44, 1.18)</td>
<td>0.19449</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>7</td>
<td>71</td>
<td>0.37 (0.16, 0.88)</td>
<td>0.02384</td>
</tr>
<tr>
<td>Male</td>
<td></td>
<td>0</td>
<td>82</td>
<td>316</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>30</td>
<td>223</td>
<td>0.53 (0.34, 0.84)</td>
<td>0.0065</td>
</tr>
</tbody>
</table>

*HWE p-value calculated from independent sample (n=550 for MI vs. n=2505 for non-MI)
<table>
<thead>
<tr>
<th>CHR</th>
<th>SNP</th>
<th>BP</th>
<th>HWE p-value</th>
<th>MAF</th>
<th>NAF</th>
<th>GWAS p-value</th>
<th>Rank</th>
<th>GWAS q-value</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>R7512462</td>
<td>204183322</td>
<td>SLC26A9</td>
<td>0.0429</td>
<td>1</td>
<td>0.0419</td>
<td>8</td>
<td>0.0007</td>
<td>25</td>
</tr>
<tr>
<td>2</td>
<td>R1476367</td>
<td>204183322</td>
<td>SLC26A9</td>
<td>0.0483</td>
<td>1</td>
<td>0.0484</td>
<td>1</td>
<td>0.0007</td>
<td>25</td>
</tr>
<tr>
<td>3</td>
<td>R7141053</td>
<td>204183322</td>
<td>SLC26A9</td>
<td>0.0462</td>
<td>2</td>
<td>0.0484</td>
<td>1</td>
<td>0.0007</td>
<td>25</td>
</tr>
<tr>
<td>4</td>
<td>R1214091</td>
<td>204183322</td>
<td>SLC26A9</td>
<td>0.0462</td>
<td>2</td>
<td>0.0484</td>
<td>1</td>
<td>0.0007</td>
<td>25</td>
</tr>
<tr>
<td>5</td>
<td>R1495465</td>
<td>204183322</td>
<td>SLC26A9</td>
<td>0.0483</td>
<td>1</td>
<td>0.0484</td>
<td>1</td>
<td>0.0007</td>
<td>25</td>
</tr>
<tr>
<td>6</td>
<td>R17583103</td>
<td>204183322</td>
<td>SLC26A9</td>
<td>0.0483</td>
<td>1</td>
<td>0.0484</td>
<td>1</td>
<td>0.0007</td>
<td>25</td>
</tr>
<tr>
<td>7</td>
<td>R17583103</td>
<td>204183322</td>
<td>SLC26A9</td>
<td>0.0483</td>
<td>1</td>
<td>0.0484</td>
<td>1</td>
<td>0.0007</td>
<td>25</td>
</tr>
<tr>
<td>8</td>
<td>R1495465</td>
<td>204183322</td>
<td>SLC26A9</td>
<td>0.0483</td>
<td>1</td>
<td>0.0484</td>
<td>1</td>
<td>0.0007</td>
<td>25</td>
</tr>
</tbody>
</table>

**Figure 25a**

*Using ALL CF patients (N=3655)*
## II. Using ΔF508 Homozygous CF patients (N=2675)

<table>
<thead>
<tr>
<th>CHR</th>
<th>SNP</th>
<th>BP</th>
<th>Gene</th>
<th>MAF</th>
<th>HWE p-value</th>
<th>GWAS Association p-value</th>
<th>GWAS</th>
<th>GWAS-HD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>rs2991956</td>
<td>232557441</td>
<td>SLC35F3</td>
<td>0.317</td>
<td>0.3105</td>
<td>2.66e-06 0.2337 4</td>
<td>0.2987</td>
<td>Rank 24</td>
</tr>
<tr>
<td>1</td>
<td>rs12410522</td>
<td>232553569</td>
<td>SLC35F3</td>
<td>0.4716</td>
<td>0.1275</td>
<td>2.78e-06 0.2337 5</td>
<td>0.2987</td>
<td>Rank 25</td>
</tr>
<tr>
<td>1</td>
<td>rs7419153</td>
<td>2041833932</td>
<td>SLC26A9</td>
<td>0.4352</td>
<td>0.2654</td>
<td>2.92e-06 0.2337 6</td>
<td>0.0033</td>
<td>Rank 2</td>
</tr>
<tr>
<td>1</td>
<td>rs7415921</td>
<td>204177506</td>
<td>SLC26A9</td>
<td>0.4293</td>
<td>0.7158</td>
<td>4.10e-06 0.2337 8</td>
<td>0.0033</td>
<td>Rank 3</td>
</tr>
<tr>
<td>1</td>
<td>rs4077468</td>
<td>204181380</td>
<td>SLC26A9</td>
<td>0.4293</td>
<td>0.4858</td>
<td>4.32e-06 0.2337 10</td>
<td>0.0033</td>
<td>Rank 4</td>
</tr>
<tr>
<td>1</td>
<td>rs4077469</td>
<td>204181508</td>
<td>SLC26A9</td>
<td>0.4293</td>
<td>0.4858</td>
<td>4.32e-06 0.2337 9</td>
<td>0.0033</td>
<td>Rank 5</td>
</tr>
<tr>
<td>1</td>
<td>rs7512462</td>
<td>204166210</td>
<td>SLC26A9</td>
<td>0.4293</td>
<td>0.7158</td>
<td>8.35e-06 0.4206 11</td>
<td>0.0061</td>
<td>Rank 6</td>
</tr>
<tr>
<td>1</td>
<td>rs12047830</td>
<td>204183322</td>
<td>SLC26A9</td>
<td>0.4905</td>
<td>0.7191</td>
<td>2.80e-05 0.7517 19</td>
<td>0.0166</td>
<td>Rank 7</td>
</tr>
<tr>
<td>1</td>
<td>rs12741299</td>
<td>204181139</td>
<td>SLC26A9</td>
<td>0.07468</td>
<td>0.7163</td>
<td>4e-04 0.8468 25</td>
<td>0.1269</td>
<td>Rank 13</td>
</tr>
<tr>
<td>2</td>
<td>rs10179921</td>
<td>43921795</td>
<td>ABCG8</td>
<td>0.0508</td>
<td>0.4074</td>
<td>3.25e-05 0.7517 23</td>
<td>0.0166</td>
<td>Rank 8</td>
</tr>
<tr>
<td>2</td>
<td>rs4953020</td>
<td>43916400</td>
<td>ABCG5</td>
<td>0.3118</td>
<td>0.3517</td>
<td>5e-04 0.8648 30</td>
<td>0.1381</td>
<td>Rank 14</td>
</tr>
<tr>
<td>3</td>
<td>rs4604609</td>
<td>10423426</td>
<td>ATP2B2</td>
<td>0.4703</td>
<td>0.9361</td>
<td>1e-04 0.8626 63</td>
<td>0.0467</td>
<td>Rank 10</td>
</tr>
<tr>
<td>3</td>
<td>rs1318189</td>
<td>10413649</td>
<td>ATP2B2</td>
<td>0.4733</td>
<td>0.6024</td>
<td>2e-04 0.8626 84</td>
<td>0.0555</td>
<td>Rank 11</td>
</tr>
<tr>
<td>3</td>
<td>rs4954353</td>
<td>10443723</td>
<td>ATP2B2</td>
<td>0.4481</td>
<td>0.2582</td>
<td>3e-04 0.8648 17</td>
<td>0.0941</td>
<td>Rank 12</td>
</tr>
<tr>
<td>5</td>
<td>rs17563161</td>
<td>550624</td>
<td>SLC9A3</td>
<td>0.2422</td>
<td>0.4466</td>
<td>6.42e-05 0.8418 40</td>
<td>0.0286</td>
<td>Rank 9</td>
</tr>
<tr>
<td>7</td>
<td>rs11976308</td>
<td>45538828</td>
<td>ADCY1</td>
<td>0.02655</td>
<td>0.6942</td>
<td>1.48e-05 0.6166 13</td>
<td>0.8837</td>
<td>Rank 1601</td>
</tr>
<tr>
<td>12</td>
<td>rs4403634</td>
<td>37837660</td>
<td>LOCE42539</td>
<td>0.3166</td>
<td>0.1522</td>
<td>1.99e-05 0.7517 14</td>
<td>0.8837</td>
<td>Rank 1620</td>
</tr>
<tr>
<td>19</td>
<td>rs2161519</td>
<td>38052643</td>
<td>SLC7A9</td>
<td>0.188</td>
<td>0.3852</td>
<td>1.08e-05 0.4867 12</td>
<td>0.8288</td>
<td>Rank 624</td>
</tr>
<tr>
<td>19</td>
<td>rs4805825</td>
<td>38106260</td>
<td>CCDC123</td>
<td>0.188</td>
<td>0.3963</td>
<td>2.34e-05 0.7517 15</td>
<td>0.8837</td>
<td>Rank 1603</td>
</tr>
<tr>
<td>X</td>
<td>rs3787866</td>
<td>115480867</td>
<td>SLC6A14</td>
<td>0.4085</td>
<td>0.9012</td>
<td>1.10e-09 6.00e-1</td>
<td>0.0006</td>
<td>Rank 1</td>
</tr>
<tr>
<td>X</td>
<td>rs5905283</td>
<td>115479909</td>
<td>SLC6A14</td>
<td>0.4974</td>
<td>0.7224</td>
<td>6.29e-07 0.1701 2</td>
<td>0.169</td>
<td>Rank 15</td>
</tr>
<tr>
<td>X</td>
<td>rs12839317</td>
<td>115479578</td>
<td>SLC6A14</td>
<td>0.249</td>
<td>0.4144</td>
<td>1.07e-06 0.1922 3</td>
<td>0.191</td>
<td>Rank 16</td>
</tr>
<tr>
<td>X</td>
<td>rs5905284</td>
<td>115485218</td>
<td>SLC6A14</td>
<td>0.2608</td>
<td>0.2653</td>
<td>3.37e-06 0.2337 7</td>
<td>0.302</td>
<td>Rank 26</td>
</tr>
</tbody>
</table>
### Figure 25c

**Additive model**

<table>
<thead>
<tr>
<th>Minor Allele</th>
<th>estimate</th>
<th>se</th>
<th>wald</th>
<th>p</th>
<th>OR(1vs0)</th>
<th>OR(2vs0)</th>
<th>95% CI of estimate</th>
<th>95% CI of OR(1vs0)</th>
<th>95% CI of OR(2vs0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs7812462_1</td>
<td>C</td>
<td>-0.3351</td>
<td>0.0732</td>
<td>19.864</td>
<td>8.32E-06</td>
<td>0.718245096</td>
<td>0.511579551</td>
<td>0.482922</td>
<td>0.197738</td>
</tr>
<tr>
<td>rs7415921_2</td>
<td>C</td>
<td>0.3369</td>
<td>0.0732</td>
<td>21.19</td>
<td>4.20E-06</td>
<td>1.400599997</td>
<td>1.96167755</td>
<td>0.193428</td>
<td>0.400372</td>
</tr>
<tr>
<td>rs4077468_1</td>
<td>C</td>
<td>-0.3421</td>
<td>0.0744</td>
<td>21.12</td>
<td>4.30E-06</td>
<td>0.710277717</td>
<td>0.504493663</td>
<td>0.487924</td>
<td>0.196276</td>
</tr>
<tr>
<td>rs4077469_4</td>
<td>T</td>
<td>-0.3421</td>
<td>0.0744</td>
<td>21.12</td>
<td>4.30E-06</td>
<td>0.710277717</td>
<td>0.504493663</td>
<td>0.487924</td>
<td>0.196276</td>
</tr>
<tr>
<td>rs12047830_2</td>
<td>C</td>
<td>0.0035</td>
<td>0.0728</td>
<td>17.4</td>
<td>3.00E-05</td>
<td>1.35491591</td>
<td>1.834918378</td>
<td>0.160812</td>
<td>0.446188</td>
</tr>
<tr>
<td>rs7419153_4</td>
<td>T</td>
<td>0.3529</td>
<td>0.0756</td>
<td>21.79</td>
<td>3.00E-05</td>
<td>1.423188817</td>
<td>2.02846641</td>
<td>0.204724</td>
<td>0.501076</td>
</tr>
<tr>
<td>rs10178921_4</td>
<td>T</td>
<td>0.5674</td>
<td>0.135</td>
<td>17.41</td>
<td>3.00E-05</td>
<td>1.756639417</td>
<td>3.085766252</td>
<td>0.5982</td>
<td>0.826</td>
</tr>
<tr>
<td>rs4650459_1</td>
<td>A</td>
<td>-0.206</td>
<td>0.074</td>
<td>14.98</td>
<td>3.00E-05</td>
<td>0.751262261</td>
<td>0.564955185</td>
<td>-0.43104</td>
<td>-0.14096</td>
</tr>
<tr>
<td>rs17563161_4</td>
<td>T</td>
<td>0.3272</td>
<td>0.0812</td>
<td>16.22</td>
<td>5.00E-05</td>
<td>1.307078565</td>
<td>1.923907779</td>
<td>0.165044</td>
<td>0.480352</td>
</tr>
</tbody>
</table>

**Genotype Model for snp on X**

<table>
<thead>
<tr>
<th>Minor Allele</th>
<th>#minor allele</th>
<th>estimate</th>
<th>se</th>
<th>wald</th>
<th>p</th>
<th>OR</th>
<th>95% CI of estimate</th>
<th>95% CI of OR</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs7887662_1</td>
<td>C</td>
<td>-0.3867</td>
<td>0.1319</td>
<td>8.99</td>
<td>0.00338</td>
<td>0.679294653</td>
<td>0.645224</td>
<td>0.128176</td>
</tr>
<tr>
<td>rs7812462_1</td>
<td>1 C</td>
<td>-0.3867</td>
<td>0.1319</td>
<td>8.99</td>
<td>0.00338</td>
<td>0.679294653</td>
<td>0.645224</td>
<td>0.128176</td>
</tr>
<tr>
<td></td>
<td>2 C</td>
<td>-0.7731</td>
<td>0.1303</td>
<td>35.16</td>
<td>3.00E-09</td>
<td>0.461579988</td>
<td>1.028488</td>
<td>0.517712</td>
</tr>
</tbody>
</table>
### Table

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>Gene Locus</th>
<th>Gene Name</th>
<th>Gene Locus</th>
<th>Gene Name</th>
<th>Gene Locus</th>
<th>Gene Name</th>
<th>Gene Locus</th>
</tr>
</thead>
<tbody>
<tr>
<td>STT3G074P01</td>
<td>T1R90167</td>
<td>SLT3G074P03</td>
<td>T1R90245</td>
<td>T1R90321P01</td>
<td>T1R90321P03</td>
<td>T1R90321P05</td>
<td>T1R90321P07</td>
</tr>
<tr>
<td>MLE1</td>
<td>T1R90167</td>
<td>MLE2</td>
<td>T1R90245</td>
<td>MLE3</td>
<td>T1R90321P01</td>
<td>MLE4</td>
<td>T1R90321P03</td>
</tr>
<tr>
<td>T1R2</td>
<td>T1R90167</td>
<td>T1R4</td>
<td>T1R90245</td>
<td>T1R6</td>
<td>T1R90321P01</td>
<td>T1R7</td>
<td>T1R90321P03</td>
</tr>
<tr>
<td>T1R1</td>
<td>T1R90167</td>
<td>T1R3</td>
<td>T1R90245</td>
<td>T1R5</td>
<td>T1R90321P01</td>
<td>T1R8</td>
<td>T1R90321P03</td>
</tr>
<tr>
<td>T1R4</td>
<td>T1R90167</td>
<td>T1R2</td>
<td>T1R90245</td>
<td>T1R3</td>
<td>T1R90321P01</td>
<td>T1R1</td>
<td>T1R90321P03</td>
</tr>
<tr>
<td>T1R7</td>
<td>T1R90167</td>
<td>T1R5</td>
<td>T1R90245</td>
<td>T1R6</td>
<td>T1R90321P01</td>
<td>T1R8</td>
<td>T1R90321P03</td>
</tr>
<tr>
<td>T1R3</td>
<td>T1R90167</td>
<td>T1R4</td>
<td>T1R90245</td>
<td>T1R2</td>
<td>T1R90321P01</td>
<td>T1R1</td>
<td>T1R90321P03</td>
</tr>
<tr>
<td>T1R8</td>
<td>T1R90167</td>
<td>T1R6</td>
<td>T1R90245</td>
<td>T1R5</td>
<td>T1R90321P01</td>
<td>T1R7</td>
<td>T1R90321P03</td>
</tr>
<tr>
<td>T1R1</td>
<td>T1R90167</td>
<td>T1R3</td>
<td>T1R90245</td>
<td>T1R4</td>
<td>T1R90321P01</td>
<td>T1R2</td>
<td>T1R90321P03</td>
</tr>
<tr>
<td>T1R2</td>
<td>T1R90167</td>
<td>T1R4</td>
<td>T1R90245</td>
<td>T1R3</td>
<td>T1R90321P01</td>
<td>T1R1</td>
<td>T1R90321P03</td>
</tr>
<tr>
<td>T1R6</td>
<td>T1R90167</td>
<td>T1R5</td>
<td>T1R90245</td>
<td>T1R2</td>
<td>T1R90321P01</td>
<td>T1R3</td>
<td>T1R90321P03</td>
</tr>
<tr>
<td>T1R5</td>
<td>T1R90167</td>
<td>T1R6</td>
<td>T1R90245</td>
<td>T1R4</td>
<td>T1R90321P01</td>
<td>T1R1</td>
<td>T1R90321P03</td>
</tr>
<tr>
<td>T1R4</td>
<td>T1R90167</td>
<td>T1R2</td>
<td>T1R90245</td>
<td>T1R3</td>
<td>T1R90321P01</td>
<td>T1R5</td>
<td>T1R90321P03</td>
</tr>
<tr>
<td>T1R3</td>
<td>T1R90167</td>
<td>T1R4</td>
<td>T1R90245</td>
<td>T1R2</td>
<td>T1R90321P01</td>
<td>T1R6</td>
<td>T1R90321P03</td>
</tr>
<tr>
<td>T1R2</td>
<td>T1R90167</td>
<td>T1R3</td>
<td>T1R90245</td>
<td>T1R4</td>
<td>T1R90321P01</td>
<td>T1R5</td>
<td>T1R90321P03</td>
</tr>
<tr>
<td>T1R1</td>
<td>T1R90167</td>
<td>T1R3</td>
<td>T1R90245</td>
<td>T1R4</td>
<td>T1R90321P01</td>
<td>T1R2</td>
<td>T1R90321P03</td>
</tr>
<tr>
<td>T1R4</td>
<td>T1R90167</td>
<td>T1R1</td>
<td>T1R90245</td>
<td>T1R3</td>
<td>T1R90321P01</td>
<td>T1R5</td>
<td>T1R90321P03</td>
</tr>
<tr>
<td>T1R3</td>
<td>T1R90167</td>
<td>T1R2</td>
<td>T1R90245</td>
<td>T1R4</td>
<td>T1R90321P01</td>
<td>T1R1</td>
<td>T1R90321P03</td>
</tr>
<tr>
<td>T1R2</td>
<td>T1R90167</td>
<td>T1R1</td>
<td>T1R90245</td>
<td>T1R3</td>
<td>T1R90321P01</td>
<td>T1R4</td>
<td>T1R90321P03</td>
</tr>
<tr>
<td>T1R1</td>
<td>T1R90167</td>
<td>T1R2</td>
<td>T1R90245</td>
<td>T1R4</td>
<td>T1R90321P01</td>
<td>T1R3</td>
<td>T1R90321P03</td>
</tr>
</tbody>
</table>

### Figure 26

![Image of Figure 26](image-url)

### Notes

- The table includes various gene names and their associated gene loci.
- Each entry in the table represents a pair of gene names with corresponding gene loci.
- The figure provides a visual representation of the data, possibly illustrating the distribution or relationship of these genes.

---

**PCT/US2011/054318**

28/55
## Figure 30

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>Sample Used</th>
<th>SNP Name</th>
<th>SNP BP</th>
<th>Association p-value</th>
<th>GWAS-HD q-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLC6A14*</td>
<td>ALL</td>
<td>rs3788766</td>
<td>115480867</td>
<td>2.09e-12</td>
<td>0</td>
</tr>
<tr>
<td>SLC26A9</td>
<td>ALL</td>
<td>rs7419153</td>
<td>204183932</td>
<td>2.81e-7</td>
<td>0.0007</td>
</tr>
<tr>
<td>PTK2</td>
<td>ALL</td>
<td>rs306962</td>
<td>142136832</td>
<td>3.00e-6</td>
<td>0.3201</td>
</tr>
<tr>
<td>PRSS1</td>
<td>ALL</td>
<td>rs1985888</td>
<td>142142146</td>
<td>3.58e-6</td>
<td>0.3201</td>
</tr>
<tr>
<td>PRKCB</td>
<td>ALL</td>
<td>rs11648535</td>
<td>23993025</td>
<td>6.48e-6</td>
<td>0.4345</td>
</tr>
<tr>
<td>SLC6A14</td>
<td>ΔF/ΔF</td>
<td>rs3788766</td>
<td>115480867</td>
<td>1.10e-9</td>
<td>0.0006</td>
</tr>
<tr>
<td>SLC26A9</td>
<td>ΔF/ΔF</td>
<td>rs7419153</td>
<td>204183932</td>
<td>2.92e-6</td>
<td>0.0038</td>
</tr>
<tr>
<td>ABCG8</td>
<td>ΔF/ΔF</td>
<td>rs10179921</td>
<td>43921795</td>
<td>3.25e-5</td>
<td>0.0166</td>
</tr>
<tr>
<td>SLC9A3</td>
<td>ΔF/ΔF</td>
<td>rs17563161</td>
<td>550624</td>
<td>6.42e-5</td>
<td>0.0286</td>
</tr>
<tr>
<td>ATP2B2</td>
<td>ΔF/ΔF</td>
<td>rs4684689</td>
<td>10423426</td>
<td>1.00e-4</td>
<td>0.0467</td>
</tr>
<tr>
<td>SLC35F3</td>
<td>ΔF/ΔF</td>
<td>rs2991956</td>
<td>232557441</td>
<td>2.66e-6</td>
<td>0.2987</td>
</tr>
</tbody>
</table>
Figure 32

CEBPB  1  TCCTTACATACCATCTACATCTTTTCTCTACCTTAAGCTC  50
         |………………………………………………………………………………
rs3788766  1  --------------------------TCTTTGATGACTTTTCCTACCTTAAGCTC  31
CEBPB  51  CAAGATTCACCAGGCAATATCCTATTGCAATTGGAATATTTGCGAGTC  100
         |………………………………………………………………………………
rs3788766  32  CAAGATTCACCAGGCAATATCCTATTGCAATTGGAATATTTGCGAGTC  81
CEBPB  101  AGTGGAGAACAAGAGAGCTATACATAGGTAGACAGATTTGGCCATTGCTT  150
         |………………………………………………………………………………
rs3788766  82  AGTGGAGAACAAGAGAGCTATACATAGGTAGACAGATTTGGCCATTGCTT  131
CEBPB  151  CATCCCTCTGTTGCAATTGTGCCTCTTCAATAGAAACCAGCATCAGCA  200
         |………………………………………………………………………………
rs3788766  132  CATCCCTCTGTTGCAATTGTGCCTCTTCAATAGAAACCAGCATCAGCA  181
CEBPB  201  CTATCTGTACCTCGCAGACGCTGGTTAAATCTTGGAGTTTCTGATG  250
         |………………………………………………………………………………
rs3788766  182  CTATCTGTACCTCGCAGACGCTGGTTAAATCTTGGAGTTTCTGATG  201
CEBPB  251  AAGAACGACATTCCATGGTAACACAGCATTTATATAATTATTTCCCTGG  300
         |………………………………………………………………………………
rs3788766  201  -----------------------------------------------  201
CEBPB  301  TCAAGCTGGATGTTCCTCATTAGTTTACTTTCT  335
         |………………………………………………………………………………
rs3788766  201  -----------------------------------------------  201
<table>
<thead>
<tr>
<th>Gene Name</th>
<th>CHR</th>
<th>SNP</th>
<th>BP</th>
<th>MAF</th>
<th>MI p-value</th>
<th>Lung p-value</th>
<th>Combined p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intergenic</td>
<td>4</td>
<td>rs4693641</td>
<td>84895294</td>
<td>0.289</td>
<td>2.22E-03</td>
<td>1.06E-03</td>
<td>6.73E-06</td>
</tr>
<tr>
<td>Intergenic</td>
<td>4</td>
<td>rs10018118</td>
<td>84969243</td>
<td>0.2735</td>
<td>7.60E-04</td>
<td>4.01E-03</td>
<td>9.12E-06</td>
</tr>
<tr>
<td>AHRR</td>
<td>5</td>
<td>rs11743226</td>
<td>457713</td>
<td>0.4164</td>
<td>4.97E-03</td>
<td>2.91E-04</td>
<td>4.79E-06</td>
</tr>
<tr>
<td>AHRR</td>
<td>5</td>
<td>rs12188164</td>
<td>481236</td>
<td>0.3771</td>
<td>6.66E-03</td>
<td>7.24E-06</td>
<td>2.97E-07</td>
</tr>
<tr>
<td>EXOC3</td>
<td>5</td>
<td>rs2672757</td>
<td>507732</td>
<td>0.4781</td>
<td>7.14E-03</td>
<td>3.30E-05</td>
<td>1.12E-06</td>
</tr>
<tr>
<td>EXOC3</td>
<td>5</td>
<td>rs2672734</td>
<td>508696</td>
<td>0.3894</td>
<td>1.64E-02</td>
<td>2.68E-05</td>
<td>2.63E-06</td>
</tr>
<tr>
<td>EXOC3</td>
<td>5</td>
<td>rs2251843</td>
<td>510955</td>
<td>0.4903</td>
<td>7.07E-03</td>
<td>2.73E-05</td>
<td>9.52E-07</td>
</tr>
<tr>
<td>SLC9A3</td>
<td>5</td>
<td>rs6868535</td>
<td>548377</td>
<td>0.4122</td>
<td>2.64E-03</td>
<td>5.41E-05</td>
<td>5.41E-07</td>
</tr>
<tr>
<td>SLC9A3</td>
<td>5</td>
<td>rs9885589</td>
<td>549577</td>
<td>0.4146</td>
<td>1.22E-03</td>
<td>5.08E-05</td>
<td>2.19E-07</td>
</tr>
<tr>
<td>SLC9A3</td>
<td>5</td>
<td>rs17563161</td>
<td>550624</td>
<td>0.2511</td>
<td>5.57E-04</td>
<td>1.52E-03</td>
<td>2.52E-06</td>
</tr>
<tr>
<td>SLC9A3</td>
<td>5</td>
<td>rs6864158</td>
<td>559051</td>
<td>0.4324</td>
<td>9.41E-05</td>
<td>2.98E-04</td>
<td>8.96E-08</td>
</tr>
<tr>
<td>TRB</td>
<td>7</td>
<td>rs1964986</td>
<td>142127214</td>
<td>0.4006</td>
<td>3.29E-05</td>
<td>2.16E-02</td>
<td>4.64E-06</td>
</tr>
</tbody>
</table>
### Figure 36

<table>
<thead>
<tr>
<th>SNP</th>
<th>CHR</th>
<th>POS</th>
<th>GENE</th>
<th>Risk Allele</th>
<th>Risk Allele Frequency</th>
<th>Discovery Sample (611:3152)(^c)</th>
<th>Replication Sample(^a)</th>
<th>N. America (251:889)(^d)</th>
<th>France (154:1078)(^d)</th>
<th>Combined(^b) (405:1967)(^d)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MI</td>
<td>MI</td>
<td>OR</td>
<td>P</td>
<td>OR</td>
<td>P</td>
<td>OR</td>
</tr>
<tr>
<td>rs4077468</td>
<td>1</td>
<td>204181380</td>
<td>SLC26A9</td>
<td>T</td>
<td>0.66</td>
<td>0.57</td>
<td>1.45</td>
<td><strong>9.88x10^{-9}</strong></td>
<td>1.45</td>
<td>0.0005</td>
</tr>
<tr>
<td>rs7512462</td>
<td>1</td>
<td>204166218</td>
<td>SLC26A9</td>
<td>T</td>
<td>0.66</td>
<td>0.57</td>
<td>1.45</td>
<td><strong>2.14x10^{-8}</strong></td>
<td>1.31</td>
<td>0.0134</td>
</tr>
<tr>
<td>rs7419153</td>
<td>1</td>
<td>204183932</td>
<td>SLC26A9</td>
<td>T</td>
<td>0.44</td>
<td>0.36</td>
<td>1.42</td>
<td><strong>1.01x10^{-7}</strong></td>
<td>1.42</td>
<td>0.0007</td>
</tr>
<tr>
<td>rs12047830</td>
<td>1</td>
<td>204183322</td>
<td>SLC26A9</td>
<td>C</td>
<td>0.56</td>
<td>0.49</td>
<td>1.34</td>
<td><strong>3.72x10^{-6}</strong></td>
<td>1.33</td>
<td>0.0054</td>
</tr>
<tr>
<td>rs3788766</td>
<td>X</td>
<td>115480867</td>
<td>SLC6A14</td>
<td>T</td>
<td>0.72</td>
<td>0.59</td>
<td>1.50</td>
<td><strong>1.28x10^{-12}</strong></td>
<td>1.14</td>
<td>0.1328</td>
</tr>
<tr>
<td>rs5995283</td>
<td>X</td>
<td>115479909</td>
<td>SLC6A14</td>
<td>C</td>
<td>0.61</td>
<td>0.50</td>
<td>1.34</td>
<td><strong>1.69x10^{-6}</strong></td>
<td>1.04</td>
<td>0.0030</td>
</tr>
<tr>
<td>rs12689137</td>
<td>X</td>
<td>115479578</td>
<td>SLC6A14</td>
<td>C</td>
<td>0.82</td>
<td>0.75</td>
<td>1.39</td>
<td><strong>1.20x10^{-6}</strong></td>
<td>1.20</td>
<td>0.0666</td>
</tr>
</tbody>
</table>

\(^a\) The North American replication collection consisted of 1,140 individuals that correspond to the continuing enrolment at the North American sites. The French cohort consisted of 1,232 patients collected at 38 French CF centers (Online Methods). The French replication samples were genotyped genome-wide with the Illumina 660W-Quad BeadChip platform while the North American samples were genotyped on the six SNPs using Taqman® Assay-on-Demand Assays.

\(^b\) The combined replication P value was calculated using the inverse-variance method, most powerful when different studies have the "same direction" of effect. The corrected type I error for the replication study is 0.008, using the conservative Bonferroni correction for the 6 SNPs tested.

\(^c\) There is another significant SNP on chromosome 1 (rs4077469, \(P = 9.88 \times 10^{-9}\)) but not included as it is separated by only 128bp and in perfect LD with rs4077468.

\(^d\) MI or Non-MI break-down for each patient category.
<table>
<thead>
<tr>
<th>Risk Allele Frequency</th>
<th>Risk Allele</th>
<th>Non-MI</th>
<th>Sample Size ($n$)</th>
<th>$P$ value</th>
<th>OR</th>
<th>OR 95%CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs3788766</td>
<td>T</td>
<td>0.737</td>
<td>816 vs. 671</td>
<td>$3.20 \times 10^{-5}$</td>
<td>1.79</td>
<td>1.36</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Female (1 vs. 2)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>255 vs. 671</td>
<td>$2.80 \times 10^{-3}$</td>
<td>1.89</td>
<td>1.25</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Male (0 vs. 2)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>814 vs. 1207</td>
<td>$2.10 \times 10^{-9}$</td>
<td>2.31</td>
<td>1.75</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>All (0 vs. 1 vs. 2)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1069 vs. 816 vs. 1878</td>
<td>$1.28 \times 10^{-12}$</td>
<td>1.49</td>
<td>1.34</td>
</tr>
<tr>
<td>rs5905283</td>
<td>C</td>
<td>0.605</td>
<td>849 vs. 488</td>
<td>$5.50 \times 10^{-4}$</td>
<td>1.66</td>
<td>1.25</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Female (1 vs. 2)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>405 vs. 488</td>
<td>$8.96 \times 10^{-3}$</td>
<td>1.61</td>
<td>1.13</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Male (0 vs. 2)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1014 vs. 1007</td>
<td>$3.80 \times 10^{-7}$</td>
<td>1.91</td>
<td>1.49</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>All (0 vs. 1 vs. 2)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1419 vs. 849 vs. 1495</td>
<td>$1.69 \times 10^{-8}$</td>
<td>1.33</td>
<td>1.21</td>
</tr>
<tr>
<td>SNP</td>
<td>CHR</td>
<td>POS</td>
<td>GENE</td>
<td>Risk Allele</td>
<td>without adjustment for CFTR</td>
<td>with adjustment for CFTR</td>
</tr>
<tr>
<td>----------</td>
<td>-----</td>
<td>-------</td>
<td>-------</td>
<td>-------------</td>
<td>-----------------------------</td>
<td>--------------------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Estimate</td>
<td>SE</td>
</tr>
<tr>
<td>rs4077468</td>
<td>1</td>
<td>204181380</td>
<td>SLC26A9</td>
<td>T</td>
<td>0.372</td>
<td>0.065</td>
</tr>
<tr>
<td>rs7512462</td>
<td>1</td>
<td>204166218</td>
<td>SLC26A9</td>
<td>T</td>
<td>0.371</td>
<td>0.066</td>
</tr>
<tr>
<td>rs7419153</td>
<td>1</td>
<td>204183932</td>
<td>SLC26A9</td>
<td>T</td>
<td>0.347</td>
<td>0.065</td>
</tr>
<tr>
<td>rs12047830</td>
<td>1</td>
<td>204183322</td>
<td>SLC26A9</td>
<td>C</td>
<td>0.293</td>
<td>0.063</td>
</tr>
<tr>
<td>rs3788766</td>
<td>X</td>
<td>115480867</td>
<td>SLC6A14</td>
<td>T</td>
<td>0.402</td>
<td>0.057</td>
</tr>
<tr>
<td>rs5905283</td>
<td>X</td>
<td>115479909</td>
<td>SLC6A14</td>
<td>C</td>
<td>0.294</td>
<td>0.052</td>
</tr>
<tr>
<td>rs12839137</td>
<td>X</td>
<td>115479578</td>
<td>SLC6A14</td>
<td>C</td>
<td>0.331</td>
<td>0.068</td>
</tr>
</tbody>
</table>
Figure 41

The number of GWAS SNPs within ±10 kb of the boundaries of indicated gene.

Permutation based P value for the gene. See Online Methods for details.

North American NACFGMC discovery cohort (n=3,763)

French replication cohort (n=1,232); For SLC6A14 and SLC26A9, the combined replication results were used when appropriate, combining the French sample with the independent North American sample (n=1,140, Table 2)

Indicating if any of the SNPs in the gene with association P value < 0.05 (or Wald statistic > 3.84 in Figures 2a and 2c) when testing the apical membrane hypothesis in the NACFGMC discovery or French samples; 1=yes, 0=no. Note that unlike the gene-based or Lasso analyses, this criterion is highly dependent on the number of SNPs per gene.

Indicating if the gene was in the multivariate model selected by Lasso in the discovery sample; 1=yes, 0=no.

CFTR showed some evidence for association with MI possibly reflecting the CF mutations that are present.

There were no GWAS SNPs passing the QC criteria in the French data.

Evidence for association of SLC26A9 to MI was apparent in the North American portion of the replication sample, but was not in the French cohort portion. It remains unclear if this difference is a reflection of chance and limited power due to smaller sample size of the French cohort, or if there are actual differences in the studied populations.

<table>
<thead>
<tr>
<th>GENE</th>
<th>CHR</th>
<th># SNPs</th>
<th>Discovery</th>
<th>Replication</th>
<th>Discovery</th>
<th>Replication</th>
<th>Lasso</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLC6A14</td>
<td>X</td>
<td>15</td>
<td>&lt;0.0001</td>
<td>&lt;0.001</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

**155 Apical Genes that were tagged by GWAS SNPs**

<table>
<thead>
<tr>
<th>Gene</th>
<th>Chr</th>
<th># SNPs</th>
<th>Discovery</th>
<th>Replication</th>
<th>Lasso</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABCA7</td>
<td>19</td>
<td>18</td>
<td>0.9832</td>
<td>0.03</td>
<td>0</td>
</tr>
<tr>
<td>ABCB4</td>
<td>7</td>
<td>43</td>
<td>0.5761</td>
<td>0.344</td>
<td>1</td>
</tr>
<tr>
<td>ABCC2</td>
<td>10</td>
<td>34</td>
<td>0.2613</td>
<td>0.835</td>
<td>0</td>
</tr>
<tr>
<td>ABCC6</td>
<td>16</td>
<td>31</td>
<td>0.0565</td>
<td>0.628</td>
<td>1</td>
</tr>
<tr>
<td>ABCG5</td>
<td>2</td>
<td>26</td>
<td>0.233</td>
<td>0.676</td>
<td>1</td>
</tr>
<tr>
<td>ABCG8</td>
<td>2</td>
<td>41</td>
<td>0.0261</td>
<td>0.23</td>
<td>1</td>
</tr>
<tr>
<td>ACY3</td>
<td>11</td>
<td>7</td>
<td>0.6527</td>
<td>0.305</td>
<td>0</td>
</tr>
<tr>
<td>ADAM17</td>
<td>2</td>
<td>9</td>
<td>0.3567</td>
<td>0.228</td>
<td>1</td>
</tr>
<tr>
<td>ADRB2</td>
<td>5</td>
<td>15</td>
<td>0.4085</td>
<td>0.304</td>
<td>0</td>
</tr>
<tr>
<td>AKAP1</td>
<td>1</td>
<td>50</td>
<td>0.788</td>
<td>0.859</td>
<td>0</td>
</tr>
<tr>
<td>AKAP7</td>
<td>6</td>
<td>29</td>
<td>0.6997</td>
<td>0.286</td>
<td>0</td>
</tr>
<tr>
<td>AKR1A1</td>
<td>1</td>
<td>5</td>
<td>0.5602</td>
<td>0.168</td>
<td>0</td>
</tr>
<tr>
<td>ANK2</td>
<td>4</td>
<td>116</td>
<td>0.6156</td>
<td>0.85</td>
<td>1</td>
</tr>
<tr>
<td>ANXA6</td>
<td>5</td>
<td>53</td>
<td>0.4416</td>
<td>0.369</td>
<td>1</td>
</tr>
<tr>
<td>AQP1</td>
<td>7</td>
<td>10</td>
<td>0.8664</td>
<td>0.489</td>
<td>0</td>
</tr>
<tr>
<td>AQP2</td>
<td>12</td>
<td>6</td>
<td>0.5037</td>
<td>0.596</td>
<td>0</td>
</tr>
<tr>
<td>AQP5</td>
<td>12</td>
<td>4</td>
<td>0.4272</td>
<td>0.402</td>
<td>0</td>
</tr>
<tr>
<td>AQP8</td>
<td>16</td>
<td>6</td>
<td>0.9199</td>
<td>0.56</td>
<td>0</td>
</tr>
<tr>
<td>ATP1B1</td>
<td>1</td>
<td>22</td>
<td>0.8975</td>
<td>0.992</td>
<td>0</td>
</tr>
<tr>
<td>ATP2B2</td>
<td>3</td>
<td>62</td>
<td>0.0006</td>
<td>0.283</td>
<td>1</td>
</tr>
<tr>
<td>Gene</td>
<td>Value1</td>
<td>Value2</td>
<td>Value3</td>
<td>Value4</td>
<td>Value5</td>
</tr>
<tr>
<td>--------------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
</tr>
<tr>
<td>ATP4A</td>
<td>19</td>
<td>11</td>
<td>0.5505</td>
<td>0.763</td>
<td>0</td>
</tr>
<tr>
<td>ATP6V0A4</td>
<td>7</td>
<td>39</td>
<td>0.8663</td>
<td>0.95</td>
<td>0</td>
</tr>
<tr>
<td>ATP6V0D1</td>
<td>16</td>
<td>4</td>
<td>0.1689</td>
<td>0.956</td>
<td>0</td>
</tr>
<tr>
<td>ATP6V0D2</td>
<td>8</td>
<td>20</td>
<td>0.1898</td>
<td>0.918</td>
<td>1</td>
</tr>
<tr>
<td>ATP6V1B1</td>
<td>2</td>
<td>13</td>
<td>0.6323</td>
<td>0.338</td>
<td>0</td>
</tr>
<tr>
<td>ATP6V1E1</td>
<td>22</td>
<td>12</td>
<td>0.8157</td>
<td>0.045</td>
<td>0</td>
</tr>
<tr>
<td>ATP80B1</td>
<td>18</td>
<td>27</td>
<td>0.0123</td>
<td>0.108</td>
<td>1</td>
</tr>
<tr>
<td>CA4</td>
<td>17</td>
<td>6</td>
<td>0.1653</td>
<td>0.267</td>
<td>0</td>
</tr>
<tr>
<td>CACNB3</td>
<td>12</td>
<td>4</td>
<td>0.8725</td>
<td>0.417</td>
<td>0</td>
</tr>
<tr>
<td>CAV1</td>
<td>7</td>
<td>15</td>
<td>0.1856</td>
<td>0.694</td>
<td>1</td>
</tr>
<tr>
<td>CD300LG</td>
<td>17</td>
<td>3</td>
<td>0.0972</td>
<td>0.611</td>
<td>0</td>
</tr>
<tr>
<td>CD44</td>
<td>11</td>
<td>53</td>
<td>0.3061</td>
<td>0.168</td>
<td>1</td>
</tr>
<tr>
<td>CDHR2</td>
<td>5</td>
<td>9</td>
<td>0.167</td>
<td>0.596</td>
<td>0</td>
</tr>
<tr>
<td>CFTP9</td>
<td>7</td>
<td>49</td>
<td>0.0053</td>
<td>0.001</td>
<td>1</td>
</tr>
<tr>
<td>CIB1</td>
<td>15</td>
<td>6</td>
<td>0.5588</td>
<td>0.607</td>
<td>0</td>
</tr>
<tr>
<td>CLCA4</td>
<td>1</td>
<td>8</td>
<td>0.3229</td>
<td>0.466</td>
<td>1</td>
</tr>
<tr>
<td>CRB1</td>
<td>1</td>
<td>27</td>
<td>0.3027</td>
<td>0.179</td>
<td>1</td>
</tr>
<tr>
<td>CRB3</td>
<td>19</td>
<td>6</td>
<td>0.3403</td>
<td>0.763</td>
<td>0</td>
</tr>
<tr>
<td>CSPG4</td>
<td>15</td>
<td>5</td>
<td>0.1625</td>
<td>0.43</td>
<td>1</td>
</tr>
<tr>
<td>CTSB</td>
<td>21</td>
<td>6</td>
<td>0.0021</td>
<td>0.201</td>
<td>1</td>
</tr>
<tr>
<td>CUBN</td>
<td>10</td>
<td>144</td>
<td>0.2361</td>
<td>0.393</td>
<td>1</td>
</tr>
<tr>
<td>DPEP1</td>
<td>16</td>
<td>2</td>
<td>0.0946</td>
<td>0.755</td>
<td>0</td>
</tr>
<tr>
<td>DPP4</td>
<td>2</td>
<td>12</td>
<td>0.3709</td>
<td>0.591</td>
<td>0</td>
</tr>
<tr>
<td>DUOX1</td>
<td>15</td>
<td>13</td>
<td>0.7363</td>
<td>0.317</td>
<td>0</td>
</tr>
<tr>
<td>DUOX2</td>
<td>15</td>
<td>6</td>
<td>0.7511</td>
<td>0.267</td>
<td>0</td>
</tr>
<tr>
<td>EFRFR</td>
<td>7</td>
<td>78</td>
<td>0.3477</td>
<td>0.041</td>
<td>1</td>
</tr>
<tr>
<td>ENPEP</td>
<td>4</td>
<td>26</td>
<td>0.5939</td>
<td>0.757</td>
<td>0</td>
</tr>
<tr>
<td>ERBB2</td>
<td>17</td>
<td>2</td>
<td>0.0701</td>
<td>0.853</td>
<td>1</td>
</tr>
<tr>
<td>ERBB3</td>
<td>12</td>
<td>4</td>
<td>0.3871</td>
<td>0.68</td>
<td>0</td>
</tr>
<tr>
<td>EZR</td>
<td>6</td>
<td>10</td>
<td>0.0317</td>
<td>0.023</td>
<td>1</td>
</tr>
<tr>
<td>F2RL2</td>
<td>5</td>
<td>19</td>
<td>0.7075</td>
<td>0.188</td>
<td>0</td>
</tr>
<tr>
<td>GIGF</td>
<td>11</td>
<td>1</td>
<td>0.4834</td>
<td>0.039</td>
<td>0</td>
</tr>
<tr>
<td>GJB6</td>
<td>13</td>
<td>6</td>
<td>0.8792</td>
<td>0.939</td>
<td>0</td>
</tr>
<tr>
<td>GNAT3</td>
<td>7</td>
<td>16</td>
<td>0.0365</td>
<td>0.957</td>
<td>1</td>
</tr>
<tr>
<td>GPR64</td>
<td>23</td>
<td>20</td>
<td>0.6066</td>
<td>0.51</td>
<td>0</td>
</tr>
<tr>
<td>IGFBP2</td>
<td>2</td>
<td>7</td>
<td>0.1872</td>
<td>0.586</td>
<td>1</td>
</tr>
<tr>
<td>IL6R</td>
<td>1</td>
<td>9</td>
<td>0.2712</td>
<td>0.031</td>
<td>0</td>
</tr>
<tr>
<td>INADL</td>
<td>1</td>
<td>100</td>
<td>0.6673</td>
<td>0.76</td>
<td>1</td>
</tr>
<tr>
<td>KCNA1</td>
<td>12</td>
<td>8</td>
<td>0.1868</td>
<td>0.638</td>
<td>1</td>
</tr>
<tr>
<td>KCNE1</td>
<td>21</td>
<td>28</td>
<td>0.9999</td>
<td>0.981</td>
<td>0</td>
</tr>
<tr>
<td>KCNMA1</td>
<td>10</td>
<td>225</td>
<td>0.0955</td>
<td>0.129</td>
<td>1</td>
</tr>
<tr>
<td>KIAA1919</td>
<td>6</td>
<td>5</td>
<td>0.5382</td>
<td>0.037</td>
<td>0</td>
</tr>
<tr>
<td>KNCN</td>
<td>1</td>
<td>4</td>
<td>0.9284</td>
<td>0.664</td>
<td>0</td>
</tr>
<tr>
<td>LCT</td>
<td>2</td>
<td>11</td>
<td>0.799</td>
<td>0.036</td>
<td>0</td>
</tr>
<tr>
<td>LHFP5</td>
<td>6</td>
<td>4</td>
<td>0.3254</td>
<td>0.551</td>
<td>0</td>
</tr>
<tr>
<td>LMO7</td>
<td>13</td>
<td>56</td>
<td>0.0267</td>
<td>0.782</td>
<td>1</td>
</tr>
<tr>
<td>LRP2</td>
<td>2</td>
<td>65</td>
<td>0.2323</td>
<td>0.964</td>
<td>1</td>
</tr>
<tr>
<td>LZTS1</td>
<td>8</td>
<td>10</td>
<td>0.1894</td>
<td>0.072</td>
<td>0</td>
</tr>
</tbody>
</table>
### Figure 41 (Continued)

<table>
<thead>
<tr>
<th>MAL</th>
<th>2</th>
<th>0.9152</th>
<th>0.869</th>
<th>0</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAL2</td>
<td>8</td>
<td>0.6686</td>
<td>0.789</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MGAM</td>
<td>7</td>
<td>0.2256</td>
<td>0.191</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>MIP</td>
<td>12</td>
<td>0.914</td>
<td>0.768</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MPDZ</td>
<td>9</td>
<td>0.788</td>
<td>0.781</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MREG</td>
<td>2</td>
<td>0.819</td>
<td>0.719</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>MUC1</td>
<td>3</td>
<td>0.3971</td>
<td>0.887</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MUC20^t</td>
<td>3</td>
<td>0.909</td>
<td>NA</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MUC3A</td>
<td>2</td>
<td>0.1207</td>
<td>0.498</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>MUC3B</td>
<td>7</td>
<td>0.4523</td>
<td>0.23</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MYO1A</td>
<td>12</td>
<td>0.338</td>
<td>0.632</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MYO7B</td>
<td>2</td>
<td>0.7972</td>
<td>0.575</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>NAALADL1</td>
<td>11</td>
<td>0.456</td>
<td>0.878</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>NPC1L1</td>
<td>7</td>
<td>0.4787</td>
<td>0.014</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>NRG1</td>
<td>8</td>
<td>0.8421</td>
<td>0.781</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>OTOA</td>
<td>16</td>
<td>0.2995</td>
<td>0.62</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>OTOG</td>
<td>11</td>
<td>0.5511</td>
<td>0.285</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>OXTR</td>
<td>3</td>
<td>0.8667</td>
<td>0.65</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>P2RY1</td>
<td>3</td>
<td>0.2739</td>
<td>0.569</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>P2RY2</td>
<td>11</td>
<td>0.5776</td>
<td>0.969</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>P2RY4</td>
<td>23</td>
<td>0.7478</td>
<td>0.535</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>P2RY6</td>
<td>11</td>
<td>0.6034</td>
<td>0.327</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PDPN</td>
<td>1</td>
<td>0.2071</td>
<td>0.244</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>PKM</td>
<td>12</td>
<td>0.2429</td>
<td>0.616</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PKHD1</td>
<td>6</td>
<td>0.0475</td>
<td>0.356</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>PLB1</td>
<td>2</td>
<td>0.8902</td>
<td>0.972</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PRKCI</td>
<td>3</td>
<td>0.639</td>
<td>0.756</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PROM1</td>
<td>4</td>
<td>0.2262</td>
<td>0.39</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>PROM2</td>
<td>2</td>
<td>0.4913</td>
<td>0.325</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PTK2</td>
<td>8</td>
<td>0.079</td>
<td>0.088</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>RAB14</td>
<td>9</td>
<td>0.3595</td>
<td>0.214</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>RAB27A</td>
<td>15</td>
<td>0.3257</td>
<td>0.583</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>RHCG</td>
<td>15</td>
<td>0.1909</td>
<td>0.581</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>S100G</td>
<td>23</td>
<td>0.3144</td>
<td>0.916</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SCNN1A</td>
<td>12</td>
<td>0.3261</td>
<td>0.288</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>SCNN1B</td>
<td>16</td>
<td>0.4818</td>
<td>0.826</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SCNN1G</td>
<td>16</td>
<td>0.8408</td>
<td>0.911</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SHROOM2</td>
<td>23</td>
<td>0.5546</td>
<td>0.178</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SHROOM3</td>
<td>4</td>
<td>0.4466</td>
<td>0.224</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SHROOM4</td>
<td>23</td>
<td>0.5186</td>
<td>0.778</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>SJ</td>
<td>3</td>
<td>0.7507</td>
<td>0.201</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SLC10A2</td>
<td>13</td>
<td>0.9229</td>
<td>0.433</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SLC11A2</td>
<td>12</td>
<td>0.3498</td>
<td>0.971</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SLC12A2</td>
<td>5</td>
<td>0.6973</td>
<td>0.905</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SLC12A3</td>
<td>16</td>
<td>0.6003</td>
<td>0.702</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SLC14A2</td>
<td>18</td>
<td>0.1743</td>
<td>0.724</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>SLC22A11</td>
<td>11</td>
<td>0.7337</td>
<td>0.533</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SLC22A12</td>
<td>11</td>
<td>0.0451</td>
<td>0.863</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Gene</td>
<td>Value1</td>
<td>Value2</td>
<td>Value3</td>
<td>Value4</td>
<td>Value5</td>
<td>Value6</td>
</tr>
<tr>
<td>----------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
</tr>
<tr>
<td>SLC22A16</td>
<td>11</td>
<td>14</td>
<td>0.9952</td>
<td>0.154</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>SLC22A4</td>
<td>5</td>
<td>47</td>
<td>0.1902</td>
<td>0.663</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SLC22A5</td>
<td>5</td>
<td>24</td>
<td>0.1561</td>
<td>0.375</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SLC23A1</td>
<td>5</td>
<td>1</td>
<td>0.2446</td>
<td>0.887</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SLC23A2</td>
<td>20</td>
<td>48</td>
<td>0.0884</td>
<td>0.218</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SLC26A3</td>
<td>7</td>
<td>19</td>
<td>0.5623</td>
<td>0.613</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SLC26A4</td>
<td>7</td>
<td>21</td>
<td>0.9859</td>
<td>0.997</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SLC26A9</td>
<td>1</td>
<td>39</td>
<td>&lt;0.0001</td>
<td>0.235</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SLC29A4</td>
<td>7</td>
<td>3</td>
<td>0.0703</td>
<td>0.754</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>SLC30A5</td>
<td>5</td>
<td>12</td>
<td>0.8699</td>
<td>0.907</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SLC34A2</td>
<td>4</td>
<td>12</td>
<td>0.5698</td>
<td>0.764</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SLC39A4</td>
<td>8</td>
<td>2</td>
<td>0.1793</td>
<td>0.492</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SLC3A2</td>
<td>11</td>
<td>7</td>
<td>0.115</td>
<td>0.978</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>SLC46A1</td>
<td>17</td>
<td>3</td>
<td>0.5825</td>
<td>0.145</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SLC4A5</td>
<td>2</td>
<td>40</td>
<td>0.5031</td>
<td>0.729</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>SLC4A7</td>
<td>3</td>
<td>22</td>
<td>0.5563</td>
<td>0.05</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>SLC5A1</td>
<td>22</td>
<td>15</td>
<td>0.8685</td>
<td>0.435</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SLC5A12</td>
<td>11</td>
<td>45</td>
<td>0.2398</td>
<td>0.081</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SLC5A8</td>
<td>12</td>
<td>24</td>
<td>0.5075</td>
<td>0.044</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SLC6A20</td>
<td>3</td>
<td>24</td>
<td>0.2724</td>
<td>0.564</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SLC7A5</td>
<td>16</td>
<td>16</td>
<td>0.6744</td>
<td>0.607</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SLC9A3</td>
<td>5</td>
<td>10</td>
<td>0.0001</td>
<td>0.017</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SLC9A3R1</td>
<td>17</td>
<td>8</td>
<td>0.5838</td>
<td>0.625</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SLC9A3R2</td>
<td>16</td>
<td>10</td>
<td>0.2069</td>
<td>0.091</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>SLC9A4</td>
<td>2</td>
<td>55</td>
<td>0.8562</td>
<td>0.012</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>STK39</td>
<td>2</td>
<td>64</td>
<td>0.9003</td>
<td>0.112</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>STX3</td>
<td>11</td>
<td>5</td>
<td>0.4133</td>
<td>0.71</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>STX4</td>
<td>16</td>
<td>2</td>
<td>0.2168</td>
<td>0.632</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>STXB3</td>
<td>1</td>
<td>12</td>
<td>0.6591</td>
<td>0.502</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TACSTD1</td>
<td>2</td>
<td>7</td>
<td>0.2121</td>
<td>0.594</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TCRG1</td>
<td>11</td>
<td>11</td>
<td>0.6047</td>
<td>0.718</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TF</td>
<td>3</td>
<td>13</td>
<td>0.7929</td>
<td>0.498</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TGFBR1</td>
<td>9</td>
<td>10</td>
<td>0.4879</td>
<td>0.746</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TLR9</td>
<td>3</td>
<td>4</td>
<td>0.0715</td>
<td>0.313</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>TRPM6</td>
<td>9</td>
<td>39</td>
<td>0.4276</td>
<td>0.723</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>TRPS5</td>
<td>7</td>
<td>6</td>
<td>0.3986</td>
<td>0.231</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>UMOD</td>
<td>16</td>
<td>14</td>
<td>0.2544</td>
<td>0.009</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>UPK3A</td>
<td>22</td>
<td>10</td>
<td>0.5462</td>
<td>0.024</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>VANGL2</td>
<td>1</td>
<td>10</td>
<td>0.5457</td>
<td>0.284</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

2 Apical genes that were NOT tagged by GWAS SNPs

<table>
<thead>
<tr>
<th>Gene</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSN</td>
<td>23</td>
</tr>
<tr>
<td>SLC34A3</td>
<td>9</td>
</tr>
</tbody>
</table>
Figure 42

Hypothesis Driven GWAS (GWAS-HD)

Biologically Relevant Hypothesis
(The Apical Membrane Hypothesis)

GO Annotation

A set of regions/genes/SNPs
(157 Apical genes)

Permutation testing

Low Priority Group
(Remaining GWAS SNPs)

High Priority Group
(3814 GWAS SNPs)

Statistical significance of the hypothesis (high priority group): multi-SNP association evidence

Re-evaluate the initial single SNP association evidence at the genome-wide level
Figure 43

<table>
<thead>
<tr>
<th>SNP</th>
<th>CH</th>
<th>BP</th>
<th>GWAS</th>
<th>GWAS-HD</th>
<th>Apical</th>
<th>GENE</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs3788766</td>
<td>X</td>
<td>115480867</td>
<td>1.28×10^{-12}</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>rs4077468</td>
<td>1</td>
<td>204181380</td>
<td>9.88×10^{-9}</td>
<td>0.0018</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>rs4077469</td>
<td>1</td>
<td>204181508</td>
<td>9.88×10^{-9}</td>
<td>0.0018</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>rs7512462</td>
<td>1</td>
<td>204166218</td>
<td>2.14×10^{-9}</td>
<td>0.0023</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>rs7419153</td>
<td>1</td>
<td>204183932</td>
<td>1.01×10^{-7}</td>
<td>0.0091</td>
<td>6</td>
<td>0.0001</td>
</tr>
<tr>
<td>rs7415921</td>
<td>1</td>
<td>204177506</td>
<td>4.49×10^{-7}</td>
<td>0.0346</td>
<td>7</td>
<td>0.0003</td>
</tr>
<tr>
<td>rs12047830</td>
<td>1</td>
<td>204183322</td>
<td>3.72×10^{-6}</td>
<td>0.167</td>
<td>12</td>
<td>0.0022</td>
</tr>
<tr>
<td>rs5905283</td>
<td>X</td>
<td>115479909</td>
<td>1.69×10^{-5}</td>
<td>0.0023</td>
<td>4</td>
<td>0.0045</td>
</tr>
<tr>
<td>rs1318819</td>
<td>3</td>
<td>10413649</td>
<td>2.11×10^{-5}</td>
<td>0.6309</td>
<td>18</td>
<td>0.0107</td>
</tr>
<tr>
<td>rs4684689</td>
<td>3</td>
<td>10423426</td>
<td>2.76×10^{-5}</td>
<td>0.6828</td>
<td>20</td>
<td>0.0123</td>
</tr>
<tr>
<td>rs495435</td>
<td>3</td>
<td>10443723</td>
<td>4.80×10^{-5}</td>
<td>0.7766</td>
<td>30</td>
<td>0.0191</td>
</tr>
<tr>
<td>rs12741299</td>
<td>1</td>
<td>204181139</td>
<td>1.23×10^{-4}</td>
<td>0.7766</td>
<td>81</td>
<td>0.0426</td>
</tr>
<tr>
<td>rs1874361</td>
<td>1</td>
<td>204174809</td>
<td>1.31×10^{-4}</td>
<td>0.7865</td>
<td>90</td>
<td>0.0426</td>
</tr>
<tr>
<td>rs17563161</td>
<td>5</td>
<td>550624</td>
<td>1.47×10^{-4}</td>
<td>0.796</td>
<td>98</td>
<td>0.0437</td>
</tr>
</tbody>
</table>

a The q value is a genome-wide adjusted P value that controls the false discovery rate.
b The rank indicates the genome-wide ordering of the SNPs based on the original association evidence alone (GWAS) or SFDR q value after incorporating the apical plasma membrane hypothesis (GWAS-HD). The GWAS and GWAS-HD rank results of all SNPs are provided in Supplementary Fig. 3.
c Apical indicates whether a gene was (= 1) or was not (= 0) on the generated apical membrane constituent list.
d The SNPs from ATP2B2 and SLC9A43 identified by GWAS-HD.