ONLINE METHODS

Human subjects. Consent was obtained from all participants of the North American (NA) Cystic Fibrosis (CF) 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 (HSC). 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. CF patients and CF-related phenotypes including lung function[5] and meconium ileus were collected by the NACFGMC (**Table 1**). The meconium ileus GWAS was restricted to 3,763 subjects with 'severe' (pancreatic exocrine insufficient) *CFTR* genotypes and of European decent. The 1,140 NA replication participants corresponded to the continuing collections at all sites (351 from TSS, 448 from UNC/CWRU and 341 from CGS) with known meconium ileus status based on previously described criteria or evidence of an abdominal scar[34].

There are 49 CF centers in France, caring for 5,000~6,000 CF patients. Prospective enrollment of CF patients was initiated in 38 centers in 2006. The French replication cohort included 1,362 phenotyped patients who were enrolled before June 2010, all of whom are >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 Illumina 610-Quad BeadChip as previously described[5]. DNA of the NA replication sample was extracted from whole blood or transformed lymphocytes and quantified with fluorimetry. Genotyping was performed with allele-specific fluorescent probes in Taqman® SNP Genotyping Assays (Custom or On-Demand;

Applied Biosystems) as recommended. DNA of the French replication sample 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 CNG, Evry, France.

Quality control (QC). The GWAS discovery subjects were first cleaned together as part of an initial QC effort by NACFGMC (see [5] for details), followed by further QC investigation for the purpose of this study. Samples with heterozygosity proportion <28%, sex incongruity, and of non-European ancestry as determined by the principal component (PC) analysis using EIGENSTRAT[35] were excluded. Using IBD estimates from PLINK[36] and PREST-plus[37] twelve cryptic full-sibs were identified and adjusted for relationship. 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 analyzed, among which only 14 individuals had genotype missing rates >1% (maximum=2.8%); 543,927 SNPs with MAF>2% were analyzed, among which 2,916 had missing rates >2% (maximum=10%). The missing rate for all 3,814 apical SNPs and 15 SNPs from *SLC6A14* was <2%.

For the NA replication sample, 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 QC, with cross-platform concordance rates >99%.

The French replication samples with missing rates >5%, sex incongruity, and pair-wise IBD estimates >40% were excluded, yielding 1,232 genotyped and phenotyped patients. SNPs present only

on the CNV370-Duo chip, with chip-wise missing rate >10% or MAF <6% were excluded. Overall, 554,792 SNPs were analyzed of which 256,756 were typed on both chips.

GO Annotation of the apical membrane constituent and nuclear envelope genes. Gene lists were generated as described in legends for Figures 2 and 3.

Imputation. Using MACH and minimac[8,9], genome-wide imputation was conducted for the 3,763 GWAS subjects. The reference sample was the 87 CEU subjects extracted from the EUR continental group of the 1000 genomes November 2010 release [38]. Imputed genotype data for 7,245,292 SNPs with MAF >2% and estimated imputation accuracy >0.3 (MACH's R-squared) were analyzed, and no new regions with genome-wide significant association evidence were identified. In *SLC26A9* and *SLC9A3*, the best imputed SNPs were only marginally more significant than any genotyped ($5.94x10^{-9}$ vs. $9.88x10^{-9}$, and $1.09x10^{-5}$ vs. $6.22x10^{-5}$, respectively), while in *SLC6A14* the minimum *P* was provided by a genotyped SNP.

Statistical methods. Generalized estimating equations (GEE[6]) was used for GWAS with an exchangeable correlation structure to account for the full-sib relationship in the data (the Geeglm function in R). Genotypes were coded additively for autosomal SNPs and chromosome X SNPs in females, and 0 and 2 were used for male chromosome X SNPs. A site covariate with four levels (**Table 1**) was included. Logistic regression in a sample of 3,199 unrelated individuals with the site covariate and the first seven PCs was also conducted, and the results were consistent with the GEE 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[36] v1.07 for autosomal SNPs and R for X chromosome SNPs).

GWAS-HD was used to accomplish two goals in parallel (**Supplementary Fig. 4**): (1) single-SNP analysis to establish the significance of individual SNPs at the genome-wide level after weighting according to a prioritization hypothesis (*e.g.* the apical hypothesis), analyzing and re-ranking all GWAS SNPs (*e.g.* 543,927 SNPs in the meconium ileus study) and (2) multi-SNP/gene analysis to test the significance of the prioritization hypothesis itself, by assessing whether the group of SNPs/genes defined by the hypothesis (*i.e.* the 3,814 SNPs annotated to 157 apical genes, **Supplementary Table 3**) collectively display significantly smaller *P* values (or larger association statistics, **Fig. 2a** and **2b**) 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 (the 3,814 apical SNPs) or a low priority group (all other 540,113 SNPs). Stratified FDR control (SFDR[16,17]) was then applied and *q* values were calculated separately in each group. Statistical significance at a given SNP was concluded if its *q* value <0.05; each SNP was re-ranked genome-wide according to its *q* value (the original GWAS *P* values were used to guide order if there were ties in *q*). This is equivalent to a weighted p-value approach[39] but with robust weights. In the meconium ileus application, the 3,814 apical SNPs were given weight ~124 and the remaining 540,113 SNPs with weight ~0.13.

The second task is to determine the statistical significance of the apical hypothesis used to prioritize the GWAS, which involves testing 3,814 SNPs simultaneously (or 3,420 SNPs in the French replication cohort). The meconium ileus phenotype was permuted (to preserve the LD pattern between SNPs) within each consortium site and independently 10,000 times (or 1,000). For each permutation sample, corresponding association analysis was performed using the GEE, and a sum of the Wald Chi-squared (1 df) association statistics of the 3,814 (or 3,420) SNPs was obtained. The empirical *P* value

was calculated as the proportion of the permutation samples whose sum statistics were larger than that in the observed sample.

The gene-based analysis is similar to the permutation test above with summation limited to the SNPs within ±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 significance level is $0.05/156 \approx 0.0003$ (**Supplementary Table 3**).

To determine the specific subset of the apical SNPs/genes contributing to meconium ileus susceptibility, a multivariate analysis using penalized logistic regression (Lasso[18]) was performed on 3,199 unrelated individuals (574 cases) extracted from the original 3,763 NA GWAS sample. The 3,814 apical SNPs and 15 SNPs within *SLC6A14* were considered in the joint analysis. After removing 93 SNPs in perfect LD with one another (r^2=1), 3,740 SNPs (and the site covariate) were included in the Lasso analysis using the glmnet package in R. The default option to standardize all predictors was turned off; the optimal value of the tuning parameter λ was chosen based on the mode of 50 λ values obtained from 50-repeated10-fold cross-validations (to maximize the deviance) . The final model included 48 SNPs spanning 36 different genes.

We used the pseudo R-squared[11] as an estimate of the phenotypic variance explained by the SNPs of interest. Calculations used the lrm function in R by regressing meconium ileus on (a) the 7 SNPs in **Table 2**, and (b) the 48 SNPs selected by Lasso in the NA sample. The calculation was done separately in the NA sample ((a)~5% and (b)~17%) and French sample ((a)~4% and (b)~8%). In the French sample, some of the 48 SNPs were not genotyped or did not pass QC and were replaced by imputed SNPs in highest LD.