Exome sequencing of extreme phenotypes identifies DCTN4 as a modifier of chronic Pseudomonas aeruginosa infection in cystic fibrosis

Mary J Emond1, Tin Louie1, Julia Emerson2,3, Wei Zhao1, Rasika A Mathias4, Michael R Knowles5, Fred A Wright6, Mark J Rieder7, Holly K Tabor8, Deborah A Nickerson7, Kathleen C Barnes4, National Heart, Lung, and Blood Institute (NHLBI) GO Exome Sequencing Project9, Lung GO9, Ronald L Gibson2,10 & Michael J Bamshad3,7,11

Exome sequencing has become a powerful and effective strategy for the discovery of genes underlying Mendelian disorders1. However, use of exome sequencing to identify variants associated with complex traits has been more challenging, partly because the sample sizes needed for adequate power may be very large2. One strategy to increase efficiency is to sequence individuals who are at both ends of a phenotype distribution (those with extreme phenotypes). Because the frequency of alleles that contribute to the trait are enriched in one or both phenotype extremes, a modest sample size can potentially be used to identify novel candidate genes and/or alleles3. As part of the National Heart, Lung, and Blood Institute (NHLBI) Exome Sequencing Project (ESP), we used an extreme phenotype study design to discover that variants in DCTN4, encoding a dynactin protein, are associated with time to first P. aeruginosa airway infection, chronic P. aeruginosa infection and mucoid P. aeruginosa in individuals with cystic fibrosis.

For unknown reasons, individuals with cystic fibrosis (MIM 219700) are at high risk for P. aeruginosa infection, and the airways of ~80% of adult individuals with cystic fibrosis are infected4. P. aeruginosa acquisition is associated with worse long-term pulmonary disease and survival5; and chronic P. aeruginosa infection is associated with reduced lung function, faster rate of lung function decline, increased exacerbation of disease and shorter median survival6,7. Accordingly, early eradication regimens for initial P. aeruginosa infection and aggressive treatment of chronic P. aeruginosa infection are standard of care8. Discovery of host factors that influence risk of P. aeruginosa airway infection could help identify mechanisms for increased susceptibility to P. aeruginosa infection in cystic fibrosis and define subpopulations for aggressive screening and therapy.

We successfully performed exome sequencing of 91 of 96 individuals selected for sequencing from the Early Pseudomonas Infection Control (EPIC)8 Observational Study and the North American cystic fibrosis Genetic Modifiers Study (GMS)9 to identify factors leading to P. aeruginosa infection in cystic fibrosis (Online Methods and Supplementary Table 1). Forty-three individuals with early age of onset of chronic P. aeruginosa infection (early P. aeruginosa extreme, all below the 10th percentile of age of onset) and the 48 oldest individuals who had not reached chronic P. aeruginosa infection (late P. aeruginosa extreme, all past the median age of onset) were sequenced (Online Methods and Supplementary Note). Percentiles were estimated from a distribution of 1,322 EPIC individuals. Successfully sequenced individuals in the early P. aeruginosa extreme sample had at least two consecutive quarters (3-month periods) with positive cultures for P. aeruginosa before 5 years of age (EPIC, n = 38) or had at least 10 consecutive years of positive cultures for P. aeruginosa, starting at age 1, and no more than one P. aeruginosa–negative culture (GMS, n = 5). The late P. aeruginosa extreme sample consisted of individuals who had never had P. aeruginosa infection by age 14 (EPIC, n = 38) and individuals who were P. aeruginosa free to age 20 or beyond (GMS, n = 10). The early P. aeruginosa extreme sample had a 400-fold higher frequency of P. aeruginosa–positive cultures compared to the late P. aeruginosa extreme sample (Supplementary Note).

Logistic regression was performed to test for association between phenotype group and variant scores collapsed by gene using the RVT1 method10. We initially included variants with an empirical minor allele frequency (MAF) of ≤0.125 in the collapsed gene scores to

References:

1Department of Biostatistics, University of Washington, Seattle, Washington, USA. 2Department of Pediatrics, University of Washington, Seattle, Washington, USA. 3Center for Clinical and Translational Medicine, Seattle Children’s Research Institute, Seattle, Washington, USA. 4Department of Medicine, School of Medicine, Johns Hopkins University, Baltimore, Maryland, USA. 5Cystic Fibrosis/Pulmonary Research and Treatment Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA. 6Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA. 7Department of Medicine, School of Medicine, Johns Hopkins University, Baltimore, Maryland, USA. 8Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA. 9Department of Genome Sciences, University of Washington, Seattle, Washington, USA. 10Truman-Katz Center for Pediatric Bioethics, Seattle Children’s Research Institute, Seattle, Washington, USA. 11Division of Pulmonary Medicine, Seattle Children’s Hospital, Seattle, Washington, USA. 12Division of Genetic Medicine, Seattle Children’s Hospital, Seattle, Washington, USA. Correspondence should be addressed to M.J.E. (emond@uw.edu) or M.J.B. (mbamshad@uw.edu).

Received 9 March; accepted 8 June; published online 8 July 2012; doi:10.1038/ng.2344
avoid eliminating causal variants enriched to high frequency and to allow for sampling variability, resulting in the identification of 11,542 genes in which at least one person had a variant and for which the distribution of variants was not collinear with the risk group variable (Supplementary Note). The model was adjusted for ancestry, using scores for three principal components from the principal-component decomposition of the exome data, and for CFTTR mutation risk group, by including a score that was based on whether individuals were in risk group 1 or not\textsuperscript{11} (Online Methods, Supplementary Table 2 and Supplementary Note).

After Bonferroni adjustment, a single gene, DCTN4 (encoding dynactin 4) on chromosome 5q33.1, was significantly associated with time to chronic \textit{P. aeruginosa} infection (naïve \(P = 2.2 \times 10^{-6}\); adjusted \(P = 0.025\); Supplementary Fig. 1). This result remained unchanged when the analysis was limited to variants with empirical MAF of \(5.05\) (Fig. 1a) and was robust across multiple analytical methods (Supplementary Figs. 2 and 3 and Supplementary Note). A resampling-based \(P\) value of \(1.5 \times 10^{-6}\) was obtained from 10 million equal parametric bootstrap trials of the RVT1 test under the null hypothesis of equal proportions of individuals with \textit{DCTN4} variants in each study arm. The resulting quantile-quantile plot showed no deviations from expected behavior that could have led to a spurious \(P\) value for \textit{DCTN4} (Supplementary Fig. 4).

Inspection of the exome sequencing data for \textit{DCTN4} revealed that 12 of the 43 individuals in the early \textit{P. aeruginosa} extreme sample had a missense variant in \textit{DCTN4}; 9 were heterozygous at position 150,097,883 (rs11954652; encoding p.Phe349Leu; MAF = 0.048 in European Americans according to the NHLBI Exome Variant Server (EVS)) and 3 at position 150,110,239 (rs35772018; encoding p.Tyr270Cys; MAF = 0.017 in European Americans according to EVS). Linkage disequilibrium (LD) between these variants was low, and both variants occurred at highly conserved sites (Genetic Evolutionary Rate Profiling scores of 4.1 and 5.4, respectively). None of the 48 individuals in the late \textit{P. aeruginosa} extreme sample had missense variants in \textit{DCTN4}.

On the basis of these findings, we screened \textit{DCTN4} by Sanger sequencing in 1,322 EPIC participants and selected for validation analysis all participants who were enrolled under the criterion of no previous \textit{P. aeruginosa}–positive cultures, excluding individuals in the exome sequencing study (Online Methods and Supplementary Note). The validation set of 696 individuals with varied \textit{CFTR} genotypes had a median of 22 quarters of \textit{P. aeruginosa} culture observations per subject (16,754 quarters total; Supplementary Table 3). Among the validation set, 78 participants were heterozygous and 9 were homozygous for the non-reference allele (C) at rs11954652, 15 were heterozygous for rs35772018 (2 of whom were not called at rs11954652) and 27 individuals were not called at either site. One individual was heterozygous for both missense variants. Three remaining individuals carried other missense variants, which were explored in secondary analyses (Supplementary Note).

We then assessed whether genotype at either rs11954652 or rs35772018 predicted age of first \textit{P. aeruginosa}–positive culture and/or age of onset of chronic \textit{P. aeruginosa} infection in the validation set, using the same definition of chronic \textit{P. aeruginosa} infection as in the exome study. Because individuals who were enrolled at older ages were selected for a \textit{P. aeruginosa}–negative history, we performed a Cox regression analysis stratified on enrollment age (individuals enrolled at later ages with a \textit{P. aeruginosa}–negative history were not at risk before enrollment), adjusting for CFTTR risk group, number of culture quarters on study and enrollment age (Online Methods and Supplementary Note).

The presence of at least one \textit{DCTN4} missense variant was significantly associated with both early age of first \textit{P. aeruginosa}–positive culture (\(P = 0.01\), hazard ratio (HR) = 1.4; Table 1 and Supplementary Fig. 5) and with early age at onset of chronic \textit{P. aeruginosa} infection (\(P = 0.004\), HR = 1.9; Fig. 1b and Table 1). The risk was highest in individuals with less selective bias toward a \textit{P. aeruginosa}–negative history (children enrolled before 1.5 years of age and 130 enrollees who participated in the study despite a history of \textit{P. aeruginosa}–positive cultures) (\(P = 0.004\), HR = 2.7; Fig. 1c and Supplementary Figs. 6 and 7). A stronger effect at younger ages is generally expected in this cohort, as individuals who never had \textit{P. aeruginosa} before older enrollment ages represent a group more highly enriched for resistance factors.

### Figure 1
Primary results for \textit{DCTN4} from exome discovery and validation phases. (a) Quantile-quantile plot of \(P\) values for a rare variant test of association between each gene and extreme \textit{P. aeruginosa} infection phenotypes. The most significant association (\(P = 2.2 \times 10^{-6}\)) was with \textit{DCTN4}, with a collapsed score based on two variants (at rs11954652 and rs35772018). (b) Kaplan-Meier curves comparing age at onset of chronic \textit{P. aeruginosa} infection by presence of \textit{DCTN4} variants among children in quintiles 2 and 3 of enrollment age among those reaching the endpoint (enrollment ages 1.6 to 6.7 years). Because of the need for analysis stratified on enrollment age, it was not possible to create a representative time-to-event curve with all individuals at once. This curve, showing the middle quintiles, is representative of the effect size over all strata combined: the HR for this subgroup is 2.3 (95% confidence interval (CI) = 1.3–4.5), similar to the estimate of 1.9 (\(P = 0.004\)) over the entire analysis set. (c) Kaplan-Meier curves comparing age at onset of chronic \textit{P. aeruginosa} infection by presence of \textit{DCTN4} variants among children who were not selected for a \textit{P. aeruginosa}–negative history in the EPIC validation sample. Blue line, children without \textit{DCTN4} variants (\(n = 246\)); red line, children with \textit{DCTN4} variants at rs11954652 and/or rs35772018 (\(n = 34\)). HR = 2.7 (1.4–5.3) with \(P = 0.004\). Comparison with b shows a larger baseline hazard for these children (both curves being steeper than in b shows the need for stratification on enrollment age when implementing the Cox model). (d) Kaplan-Meier curves comparing age at onset of chronic \textit{P. aeruginosa} infection among all enrollment strata by \textit{DCTN4} variant group. Blue line, no \textit{DCTN4} variants (\(n = 565\)); red line, rs11954652[C] heterozygotes (het, \(n = 78\); green line, rs11954652[C] homozygotes (hom) and rs35772018[C] heterozygotes combined (\(n = 22\)). Individuals in the latter group have higher risk than those in either of the other two groups (HR = 3.3, \(P = 0.002\) compared to baseline). Differences between groups appear somewhat compressed relative to the Cox model HR estimates because all enrollment strata are shown together: there are too few individuals in the third group to visualize differences within strata, but it is notable that a strong difference can be seen even without stratification.

### Table 1
| Genotype                | HR (95% CI) |
|-------------------------|-------------|
| rs11954652[C] hom or rs35772018[C] het | 2.7 (1.4–5.3) |

### Supplementary Table 2

| Variant       | MAF         |
|---------------|-------------|
| rs11954652    | 0.017       |
| rs35772018    | 0.017       |
| rs11954652[C] | 0.048       |
| rs35772018[C]| 0.017       |
No significant interaction was found between CFTR genotypes and DCTN4 mutations, although power to detect such an interaction was low (Supplementary Note).

Because rs11954652[C] is common in African Americans (MAF = 0.53 according to EVS), we repeated the analysis using only self-identified European Americans (n = 645) with a similar result (P = 0.03, HR = 1.7) (Supplementary Figs. 8 and 9). Additionally, for 530 individuals for whom genome-wide genotyping data were available, we repeated the analysis using principal-component decomposition to exclude all those with non-European ancestry (Supplementary Figs. 10 and 11), again with the same result (P = 0.004, HR = 2.2). These analyses confirm that effects from genetic ancestry did not confound our results. We also found a relationship between risk for P. aeruginosa infection and burden or rarity/conservation of variants. When individuals with DCTN4 mutations were subdivided by burden and the specific site of mutation, those who were homozygous for the more common variant (n = 9) or heterozygous for the rarer and more highly conserved site (n = 13) had the higher risk for early-onset P. aeruginosa infection compared to individuals with either variant (HR = 3.3, P = 0.002; Fig. 1d and Table 1) and compared to individuals heterozygous for the more common variant (HR = 2.7, P = 0.01). When individuals with the rarer rs35772018 variant were analyzed as a group by themselves, the HR for age of onset of chronic P. aeruginosa infection associated with rs35772018 genotype was estimated to be 15.9 at birth (P = 0.002) relative to individuals with neither variant (Supplementary Figs. 12–14 and Supplementary Tables 4 and 5).

Notably, estimates from analysis using evidence of 50% of cultures that were positive for P. aeruginosa infection in any 1-year period to define chronic P. aeruginosa infection (not necessarily consecutive positive cultures), analogous to the widely accepted Leeds criterion13, were similar but more significant (Supplementary Figs. 6, 7, 9, 13 and 14, Supplementary Table 5 and Supplementary Note.)

Transformation of P. aeruginosa to the mucoid state is an adaptation of P. aeruginosa to the host environment and is associated with rapid progression of cystic fibrosis airway disease14. Carriage of either DCTN4 variant was also significantly associated with age of onset of the first mucoid P. aeruginosa culture, with the effect again stronger in younger enrollees, with an estimated HR = 2.6 at birth (P = 0.03; Table 1 and Supplementary Fig. 15). Additionally, the interval between first P. aeruginosa–positive culture and first mucoid P. aeruginosa culture was significantly shorter in individuals with DCTN4 variants (HR = 3.8, P = 0.01). Analysis of a second auxiliary outcome, frequency of P. aeruginosa–positive cultures over the study period, showed that individuals with DCTN4 missense variants had higher rates of P. aeruginosa positivity (P = 0.02) (Supplementary Fig. 16). Because P. aeruginosa induces autophagy in alveolar macrophages in vitro, and autophagy has an essential role in the clearance of P. aeruginosa15. In cystic fibrosis, intracellular accumulation of CFTRAF508 is associated with reduced macroautophagic flux via inhibition of autophagosome formation. This results in increased airway inflammation15. It is possible that isoforms of dynactin 4 influence P. aeruginosa infection in cystic fibrosis by reducing autophagic clearance of P. aeruginosa in the airway of individuals with cystic fibrosis or by altering macroautophagic clearance of class II mutant CFTR (for example, ΔF508), leading to increased airway disease.

To our knowledge, this is the first study to discover a gene for a complex trait, or at minimum a genetic modifier of a Mendelian trait, using exome sequencing and an extreme phenotype study design. Notably, given the sample size of EPIC and the low frequency of rs11954652 and rs35772018 alleles, neither would have achieved genome-wide significance via genome-wide association study (GWAS), and neither variant is on common SNP-genotyping platforms or is tagged well (Supplementary Note). Our success with exome sequencing was due, in part, to the synergy of several key factors, including the use of phenotypically well-matched extremes with the exception of the trait of interest, a relatively large estimated effect size for DCTN4 and a reasonably high collective MAF for implicated variants (0.065), with the higher MAF variants intentionally included in our analyses. In most cases, use of a similar strategy to find variants underlying complex traits will likely require exome sequencing of larger sample sizes. However, we think that enthusiasm for this approach should continue as the cost of sequencing is dropping rapidly and more efficient statistical approaches for analysis of rare variants are becoming available.

URLs. Exome Variant Server (EVS), http://evs.gs.washington.edu/EVS/; OMIM, http://www.omim.org/; database of Genotypes and Phenotypes (dbGaP), http://www.ncbi.nlm.nih.gov/gap.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Exome data are available at the NCBI dbGaP repository under accession phs000254.v1.p1.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

We thank the families for their participation and the EPIC study site investigators and research coordinators (Supplementary Note) for their assistance. We thank

Table 1. Results of association analyses of DCTN4 and P. aeruginosa phenotypes

| Event | Group | N | Number of events | HR | 95% CI | P value |
|-------|-------|---|------------------|----|-------|---------|
| Age at first P. aeruginosa–positive culture | No missense | 565 | 345 | 1.0 | – | – |
| Age of onset of chronic P. aeruginosa | Any missense | 102 | 70 | 1.4 | (1.1–1.8) | 0.01 |
| Age of onset of chronic P. aeruginosa | No missense | 565 | 93 | 1.0 | – | – |
| Age of onset of chronic P. aeruginosa | Het rs11954652(C) | 78 | 20 | 1.7 | (1.0–2.7) | 0.05 |
| Age of onset of chronic P. aeruginosa | Het rs35772018(C) or hom rs11954652(C) | 22 | 8 | 3.3 | (1.5–6.9) | 0.002 |
| Age of onset of mucoid P. aeruginosa | No missense | 565 | 84 | 1.0 | – | – |
| Time from first detection of P. aeruginosa to mucoid P. aeruginosa | No missense | 410 | 55 | 1.0 | – | – |
| Time from first detection of P. aeruginosa to mucoid P. aeruginosa | Any missense | 89 | 17 | 3.8 | (1.4–10.5) | 0.01 |
Exome sequencing as a tool for Mendelian disease gene discovery. Nat. Rev. Genet. 12, 745–755 (2011).

1. Bamshad, M. J. et al. Exome sequencing as a tool for Mendelian disease gene discovery. Nat. Rev. Genet. 12, 745–755 (2011).

2. Stützle, N. O., Kiezun, A. & Sunyaev, S. Computational and statistical approaches to analyzing variants identified by exome sequencing. Genome Biol. 12, 227 (2011).

3. Lanktree, M. B., Hegele, R. A., Schork, N. J. & Spence, J. D. Extremes of unexplained variation as a phenotype: an efficient approach for genome-wide association studies of cardiovascular disease. Circ. Cardiovasc. Genet. 3, 215–221 (2010).

4. Gibson, R. L., Burns, J. L. & Ramsey, B. W. Pathophysiology and management of pulmonary infections in cystic fibrosis. Am. J. Respir. Crit. Care Med. 168, 918–951 (2003).

5. Emerson, J., Rosenfeld, M., McNamara, S., Ramsey, B. & Gibson, R. L. Pseudomonas aeruginosa and other predictors of mortality and morbidity in young children with cystic fibrosis. Pediatr. Pulmonol. 34, 91–100 (2002).

6. Proesmans, M. et al. Evaluating the “Leeds criteria” for Pseudomonas aeruginosa infection in a cystic fibrosis centre. Eur. Respir. J. 27, 937–943 (2006).

7. Johansen, H. K. et al. Antibody response to Pseudomonas aeruginosa in cystic fibrosis patients: a marker of therapeutic success?—a 30-year cohort study of survival in Danish CF patients after onset of chronic P. aeruginosa lung infection. Pediatr. Pulmonol. 37, 427–432 (2004).

8. Treggiari, M. M. et al. Early anti-pseudomonal acquisition in young patients with cystic fibrosis: rationale and design of the EPIC clinical trial and observational study. Contemp. Clin. Trials 30, 256–268 (2009).

9. Wright, F. A. et al. Genome-wide association and linkage identify modifier loci of lung disease severity in cystic fibrosis at 11p13 and 20q13.2. Nat. Genet. 43, 539–546 (2011).

10. Morris, A. P. & Zeggini, E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet. Epidemiol. 34, 188–193 (2010).

11. Green, D. M. et al. Mutations that permit residual CFTR function delay acquisition of multiple respiratory pathogens in CF patients. Respir. Res. 11, 140 (2010).

12. Lee, T. W., Brownlee, K. G., Conway, S. P., Denton, M. & Littlewood, J. M. Evaluation of a new definition for chronic Pseudomonas aeruginosa infection in cystic fibrosis patients. J. Cyst. Fibros. 2, 29–34 (2003).

13. Li, Z. et al. Longitudinal development of mucoid Pseudomonas aeruginosa infection and lung disease progression in children with cystic fibrosis. J. Am. Med. Assoc. 293, 581–588 (2005).

14. Kimura, S., Noda, T. & Yoshimori, T. Dynein-dependent movement of autophagosomes mediates efficient encounters with lysosomes. Cell Struct. Funct. 33, 109–122 (2008).

15. Haspel, J. A. & Choi, A. M. Autophagy: a core cellular process with emerging links to pulmonary disease. Am. J. Respir. Crit. Care Med. 184, 1237–1246 (2011).

16. Yuan, K. et al. Autophagy plays an essential role in the clearance of Pseudomonas aeruginosa by alveolar macrophages. J. Cell Sci. 125, 507–515 (2012).
**Online Methods**

**Exome sequencing.** Quality control of sample DNA. The initial quality control that was performed on all samples included sample quantification (PicoGreen), confirmation of high-molecular-weight DNA, test PCR amplification (four amplicons) and sex determination using a TaqMan assay. All samples were genotyped (Illumina BeadXpress) for 96 high-frequency (30–50% MAF) exome-specific SNPs, derived from the content found on genotyping chips from Illumina and Affymetrix and used to ensure sample tracking integrity throughout sample preparation and the sequencing pipeline.

**Library production and exome capture.** Approximately 3.5 µg of genomic DNA was used for a series of shotgun library construction steps, including acoustic fragmentation (Covaris), end-polishing and A-tailing, ligation of sequencing adaptors and PCR amplification. Sample shotgun libraries were captured for exome enrichment using one of three in-solution capture products: CCDS 2008 (~26 Mb), Roche/Nimblegen SeqCap EZ Human Exome Library v1.0 (~32 Mb; Roche Nimblegen EZ Cap v1) or EZ Cap v2 (~34 Mb) per the manufacturer’s instructions.

**Clustering and sequencing.** Library concentration and flow-cell loading cluster densities were determined using a standardized quantitative PCR (qPCR) protocol (Kapa Biosystems). Using the automated Illumina cBot cluster station, non-plexed samples were processed in batches of eight (one for each lane of the flow cell) and were diluted and denatured to their final effective loading concentrations. Hybridization was followed by cluster generation via bridge PCR according to standard protocols (Illumina). Enriched libraries were sequenced on an Illumina Genome Analyzer IIx using paired-end 76-base runs.

**Read mapping and variant analysis.** Samples were processed in real-time base-calls (RTA 1.7 software (Bustard)), converted to .txt files and aligned to a human reference genome (hg19) using Burrows-Wheeler Aligner (BWA). Read pairs not mapping within 2 s.d. of the average library size (~125 ± 15 bp for exomes) were removed. Data were processed using the Genome Analysis Toolkit (GATK ref.1.2005)59. All aligned read data were subjected to removal of reads with duplicate start positions, indel realignment and base-quality recalibration. Variant detection and genotyping were performed using the UnifiedGenotyper tool from GATK and were only performed on the targeted exome regions. Variant data for each sample were formatted (variant call format [VCF]) as ‘raw’ calls for all samples, and sites were flagged if they had lower quality or were false positives (those with low quality scores (≥50), allelic imbalance (≥85), long homopolymer runs (>3) and/or low quality by depth (<5)). Samples were considered complete when exome-targeted read coverage was >8× over >90% of the exome target. Typical mean coverage of the target was 60–80×.

**Data analysis quality control.** Individual exome sequencing data were evaluated against quality control metrics (Supplementary Table 1), including assessment of (i) total reads: a minimum of 30 million paired-end reads; (ii) library complexity: the ratio of unique reads to total reads mapped to the target; (iii) capture efficiency: the ratio of reads mapped to the target versus the reads mapped to the human reference; (iv) coverage distribution: 90% at ≥28× required for completion; (v) capture uniformity; (vi) raw error rates; (vii) transition/transversion (Ti/Tv) ratio (3.2 for known sites and 2.9 for novel sites); (viii) distribution of known and novel variants relative to dbSNP; (ix) fingerprint concordance of >99%; (x) homozygosity; and (xi) heterozygosity. All quality control metrics for both single-lane and merged data were reviewed to identify data deviations from known or historical norms. Lanes with values that fell outside acceptance criteria were removed. Lanes were considered complete when exome-targeted read coverage was >8× over ≥90% of the exome target. Typical mean coverage of the target was 60–80×.

**Selection of phenotypic extremes.** The definition of chronic P. aeruginosa infection—two consecutive 3-month periods with a P. aeruginosa–positive culture within each period—was chosen to be concordant with the definition used in the EPIC clinical trial to mark P. aeruginosa not eradicated by treatment.6 We determined that individuals who had reached the chronic endpoint if they had a 6-month period of positive cultures. This definition is also very similar to the Leeds criterion. Individuals in the analysis set had a median of 3.5 culture quarters per year.

**Selection of extreme individuals.** Of the 38 EPIC individuals with early-onset chronic P. aeruginosa infection in the exome sequencing sample, half were in the earliest 5% overall (≤2.5 years of age at onset), and all were in the earliest 7% (≤5 years of age) (as estimated by Kaplan-Meier curve using the 1,322 participants from the EPIC DNA Collection study). These individuals represent the extreme in terms of overall frequency of P. aeruginosa–positive quarters, with all being in the worst 20% and half among the worst 4% (>24% of quarters positive). These 38 individuals had a total of 1,165 quarters of observation time, with 334 of these quarters positive for P. aeruginosa (28%). The five GMS individuals selected for the early-onset P. aeruginosa extreme phenotype had contiguous years with positive cultures from the first or second year of life until the end of follow-up analysis (minimum follow-up age = 12, maximum = 19), with the exception of two individuals, who each had 1 year without positive cultures.

EPIC individuals for the late-onset extreme (n = 38 successful exomes) were selected from among the oldest individuals who were still P. aeruginosa free at the time of selection of the exome sequencing sample. A sample of individuals who had never had P. aeruginosa infection was chosen to balance the early-onset chronic sample in terms of sex and CFTR clinical risk group (1, 2 or unknown), although most individuals were in risk group 1 (35 in the early-onset extreme and 37 in the late-onset extreme for successful exomes). The ten individuals selected for the late-onset extreme from GMS were P. aeruginosa free until at least age 20, with one individual P. aeruginosa free until age 58. All GMS participants were in CFTR risk group 1.

**Two-sample test by gene for exome data.** A previously described method was used to obtain P values for each gene. We adjusted each gene test for CFTR risk group (risk group 1 or not) and for PC1, PC2 and PC18 from a principal-component decomposition of the entire set of exome data, after applying quality control filters (Supplementary Table 2). Indicators for sex and siblings were initially entered but were discarded after the empirical distributions of covariates for these variables were found to center at zero. The first two principal components were included in the regression model to adjust for possible ancestry-related stratification. PC18 was added to the model after examining box plots of PC3 through PC20 for cases versus controls, as PC18 showed a relatively large difference between the two groups.

**Resampling-based estimation of P value.** The nominal P value for DCTN4 from the RV1T test, as well as the general performance of the test for the specific data and sample size, was assessed via a parametric bootstrap. The parametric bootstrap was performed by drawing non-reference alleles for each person according to the joint binomial distributions for the two variants, on the basis of ESP-wide observed MAFs and correlation (rs11954652: MAF = 0.048; rs35772018: MAF = 0.0178). Alleles were drawn independently for each person under the null hypothesis of no association with phenotype group. The RV1T test was then performed using these random variants to form gene scores, while keeping each person’s covariates and phenotype group fixed. We performed 10 million trials of this bootstrap procedure to obtain a precise estimate of the P value and to examine the behavior of the RV1T test statistics in the tail of the distribution (Supplementary Fig. 4).

**Cox model analyses.** Censored data methods (Cox model and Kaplan-Meier survival curves) were used to determine that hazard ratio for time to event for participants with either of the DCTN4 variants versus those without either. Events studied were age at onset of chronic P. aeruginosa infection, age at onset of mucoid P. aeruginosa and time between first P. aeruginosa and first mucoid P. aeruginosa culture (excluding zero times). Additional analyses were carried out using separate indicators for participants who were heterozygous at the rs11954652 locus, homozygous at the rs11954652 locus and heterozygous at the rs35772018 locus. Because of the small numbers of events in the latter two groups.
(four progressions to chronic *P. aeruginosa* infection in each), these last two groups were combined to gain more precision in the estimate of effect size. All models were stratified on five groups by enrollment age (**Supplementary Note**), and all models included enrollment age—an indicator for CFTR risk group 2—and the number of observations on study, to adjust for less sensitivity for endpoint detection in individuals with fewer observations. Neither sex nor a homozygous genotype for CFTRΔ508 were significant predictors of any events (nor were their interactions with DCTN4 variant score) and were not included in the final models. Self-declared ancestry was not included in the model because there were not enough individuals of non-European ancestry to attain convergence of the estimation algorithm. However, the model was fitted for only individuals of European ancestry to determine whether the effects seen for DCTN4 were driven by the few African-American individuals in the analysis, and principal-component adjustment for ancestry was performed on the subset of 530 individuals with enough DNA for exome chip genotyping (**Supplementary Note**).

17. Yu, L., Martinez, F.D. & Klimecki, W.T. Automated high-throughput sex-typing assay. *Biotechniques* **37**, 662–664 (2004).
18. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
19. DePristo, M.A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).