HHS Public Access

Author manuscript

Nat Genet. Author manuscript; available in PMC 2013 July 05.

Published in final edited form as:

Nat Genet. ; 44(8): 886–889. doi:10.1038/ng.2344.

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

Mary J Emond[1,†], **Tin Louie**[1], **Julia Emerson**[2,3], **Wei Zhao**[1], **Rasika A. Mathias**[4], **Michael R. Knowles**[5], **Fred A. Wright**[6], **Mark J. Rieder**[7], **Holly K. Tabor**[2,8], **Debbie A. Nickerson**[7], **Kathleen C. Barnes**[4], **NHLBI GO Exome Sequencing Project**[9], **Lung GO**[9], **Ronald L. Gibson**[2,10], and **Michael J. Bamshad**[2,7,11,†]

[1]Department of Biostatistics, University of Washington, Seattle, Washington, USA

[2]Department of Pediatrics, University of Washington, Seattle, Washington, USA

[3]Center for Clinical and Translational Medicine, Seattle Children's Research Institute Seattle, Washington, USA

[4]Department of Medicine, School of Medicine, Johns Hopkins University, Baltimore, Maryland, USA

[5]Cystic Fibrosis/Pulmonary Research and Treatment Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

[6]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

[7]Department of Genome Sciences, University of Washington, Seattle, Washington, USA

[8]Trueman-Katz Center for Pediatric Bioethics, Seattle Children's Research Institute, Seattle, Washington, USA

[9]A full list of members and affiliations is provided in the Supplementary Note

[†]Corresponding authors: Mary J. Emond, PhD, Department of Biostatistics, University of Washington, Box 357232, Seattle, WA 98195, Phone: 206-543-3406, emond@u.washington.edu, Michael J. Bamshad, MD, Department of Pediatrics, University of Washington School of Medicine, Box 356320, 1959 NE Pacific Street, Seattle, WA 98195, mbamshad@u.washington.edu.

URLs

Exome Variant Server: http://evs.gs.washington.edu/EVS/
OMIM: http://www.omim.org/
dbGaP: http://www.ncbi.nlm.nih.gov/gap

Data access

Exome data are available via the NCBI dbGaP repository(Accession: phs000254.v1.p1).

Author contributions

The project was conceived and experiments planned by M.J.B., M.J.E., M.R.K., K.C.B. and R.L.G. Review of phenotypes and sample collection were performed by M.J.B., M.J.E., R.L.G., J.E, and M.R.K. Experiments were performed by M.J.R. and D.A.N. Regulatory review and guidance was provided by H.K.T. Data analysis was performed by M.J.E., J.E., T.L., F.A.W., W.Z. and R.A.M. The manuscript was written by M.J.B., M.J.E., and R.L.G. All aspects of the study were supervised by M.J.B., M.J. E., and R.L.G.

Competing Interests Statement

The authors declare no competing financial interests.

[10]Division of Pulmonary Medicine, Seattle Children's Hospital, Seattle, Washington, USA

[11]Division of Genetic Medicine, Seattle Children's Hospital, Seattle, Washington, USA

Abstract

Exome sequencing has become a powerful and effective strategy for discovery of genes underlying Mendelian disorders[1]. However, use of exome sequencing to identify variants associated with complex traits has been more challenging, partly because the samples sizes needed for adequate power may be very large[2]. One strategy to increase efficiency is to sequence individuals who are at both ends of a phenotype distribution (i.e., extreme phenotypes). Because the frequency of alleles that contribute to the trait are enriched in one or both extremes of phenotype, a modest sample size can potentially identify novel candidate genes/alleles[3]. As part of the National Heart, Lung, and Blood Institute Exome Sequencing Project (ESP), we used an extreme phenotype design to discover that variants in *DCTN4*, encoding a dynactin protein, are associated with time to first *Pseudomonas aeruginosa* (*P. aeruginosa*) airway infection, chronic *P. aeruginosa* infection and mucoid *P. aeruginosa* among individuals with cystic fibrosis (MIM219700).

For unknown reasons, individuals with cystic fibrosis are at high risk for *P. aeruginosa* infection, and the airways of ~80% of adult patients are infected[4]. *P. aeruginosa* acquisition is associated with worse long-term pulmonary disease and survival[5], and chronic *P. aeruginosa* infection is associated with reduced lung function, faster rate of lung function decline, increased exacerbations of disease, and shorter median survival[6,7]. Accordingly, early eradication regimens for initial *P. aeruginosa* infection and aggressive treatment of chronic *P. aeruginosa* infection is standard of care[8]. Discovery of host factors that influence risk of *P. aeruginosa* airway infection could help identify mechanisms for the increased susceptibility to *P. aeruginosa* infection in cystic fibrosis and define sub-populations for aggressive screening and therapy.

We performed exome sequencing successfully on 91 of 96 individuals selected for sequencing from the Early *Pseudomonas* Infection Control[8] (EPIC) 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* (i.e., early *P. aeruginosa* extreme, all within the 10th%ile of age-of-onset) and the 48 oldest individuals who had not yet reached chronic *P. aeruginosa* (i.e., 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 1322 EPIC individuals. Successfully sequenced individuals in the early *P. aeruginosa* extreme had at least two consecutive quarters (3-month periods) with positive *P. aeruginosa* cultures prior to age 5 years (EPIC, n=38) or had at least 10 years of positive *P. aeruginosa* cultures starting at age 1 and no more than one *Pa*-negative year (GMS, n=5). The late *P. aeruginosa* extreme consisted of individuals who had never had *P. aeruginosa* by age 14 (EPIC, n=38) and individuals who were *P. aeruginosa* free through age 20 or beyond (GMS, n=10). The early *P. aeruginosa*

extreme had a 400-fold higher frequency of *P. aeruginosa* positive cultures compared to the late *P. aeruginosa* extreme (Supplementary Note).

Logistic regression was performed to test for association between phenotype group and variant scores collapsed by gene as described in Morris and Zeggini (MZ)[10]. We initially included variants with an empirical MAF ≤ 0.125 in the collapsed gene scores to avoid eliminating causal variants enriched to high frequency and to allow for sampling variability, resulting in 11,542 genes for 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 (PCs) from the PC decomposition of the exome data, and for *CFTR* mutation risk group by including a score for risk group 1 or not [11] (Online Methods, Supplementary Note and Supplementary Table 2).

After Bonferroni adjustment, a single gene, *dynactin 4* (*DCTN4*) on chromosome 5q33.1, was significantly associated with time to chronic *P. aeruginosa* (naïve $p=2.2x10^{-6}$; adjusted p = 0.025; Supplementary Fig. 1). This result remained unchanged when the analysis was limited to variants with empirical MAF ≤ 0.05 (Fig. 1A) and was robust across multiple analytical methods (Supplementary Note and Supplementary Figs. 2 and 3). A resampling-based p-value of $p=1.5\times10^{-6}$ was obtained from 10 million parametric bootstrap trials of MZ under the null hypothesis. The resulting QQ-plot showed no deviations from expected behavior that could lead to a spurious p-value for *DCTN4* (Supplementary Fig. 4).

Inspection of the exome seqeunce data for *DCTN4* revealed that 12/43 individuals in the early *P. aeruginosa* extreme had a missense variant in *DCTN4*, 9 were heterozygous at position 150097883 (rs11954652; Phe349Leu; MAF=0.048 in European Americans (EA) per the NHLBI-Exome Variant Server (EVS)) and 3 at position 150110239 (rs35772018; Tyr270Cys; MAF=0.017 in EA per the EVS). LD between these variants was low and both variants occur at highly conserved sites (Genetic Evolutionary Rate Profiling scores 4.1 and 5.4, respectively). None of the 48 individuals in the late *P. aeruginosa* extreme had missense variants in *DCTN4*.

Based on these findings, we screened *DCTN4* by Sanger sequencing in 1322 EPIC participants and selected for a validation analysis all participants who were enrolled under the criterion of no prior positive *P. aeruginosa* cultures, excluding individuals in the exome sequencing study (Online Methods and Supplementary Note). The validation set of 696 individuals with varied *CFTR* genotypes had a median of 22 quarters of *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 (i.e., C) at rs11954652; 15 were heterozygous for rs35772018 (two of which were not called at rs11954652; and 27 individuals were not called at either site. One individual was heterozygous for both missense variants. Three additional individuals carried other missense variants, which were explored in secondary analyses (Supplementary Note).

We then assessed whether either rs11954652 or rs35772018 predicted age of first *P. aeruginosa* positive culture and/or age-of-onset of chronic *P. aeruginosa* in the validation

set, using the same definition of chronic *P. aeruginosa* as used in the exome study. Because individuals who were enrolled at older ages were selected for a negative *P. aeruginosa* history, we performed a Cox regression analysis stratified on enrollment age (individuals enrolled at later ages with a negative *P. aeruginosa* history are not at risk prior to enrollment), adjusting for *CFTR* risk group, number of culture-quarters on study and enrollment-age (Online Methods and Supplementary Note).

The presence of at least one *DCTN4* missense variant was significantly associated with both early age of first *P. aeruginosa* positive culture (p=0.01, HR=1.4; Table 1; Supplementary Fig. 5) and early age-at-onset of chronic *P. aeruginosa* (p=0.004, HR=1.9; Table 1; Fig. 1B). The risk was strongest in individuals with less selective bias toward a negative *P. aeruginosa* history (children enrolled before age 1.5 years of age and 103 enrollees who participated in the trial despite a positive *P. aeruginosa* culture history)(p=0.004, HR=2.7) (Fig. 1C) (Supplementary Figs. 6 and 7). A stronger effect at younger ages is not surprising in this cohort, as individuals who never had *P. aeruginosa* prior to older enrollment ages represent a group more highly enriched for resistance factors. No significant interactions were found between *CFTR* genotypes and *DCTN4* mutations, although power to detect such an interaction is low (Supplementary Note).

Since rs11954652 is common in African Americans (MAF=0.53 per the 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 results confirm that genetic ancestry did not confound our results. We also found a relationship between risk for *P. aeruginosa* and burden or rarity/conservation of variants. When individuals with *DCTN4* mutations were subdivided by burden and specific site of mutation, those who were homozygous for the more common variant (n=9) or heterozygous for the rarer and more highly conserved variant (n= 13) had higher risk for early-onset *P. aeruginosa* compared to individuals with neither variant (HR = 3.3, p=0.002; Table 1, Fig. 1D) 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 hazard ratio for age-of-onset of chronic *P. aeruginosa* associated with rs35772018was estimated to be 15.9 at birth (p=0.002) compared to individuals with neither variant (Supplementary Tables 4 and 5; and Supplementary Figs. 12, 13, and 14). Importantly, estimates from analysis using existence of 50% of *P. aeruginosa* cultures positive in any 1-year period to define chronic *P. aeruginosa* (i.e. not necessarily consecutive positive cultures) analogous to the widely accepted Leed's criterion[12], were similar but more significant (Supplementary Note, Table 5 and Supplementary Figs. 6, 7, 9, 13, and 14.)

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 disease[13]. Carriage of either *DCTN4* variant also was significantly associated with age-of-onset of first mucoid *P. aeruginosa* culture, the effect again stronger in younger enrollees

with an estimated HR = 2.6at birth (p=0.03; Table 1; Supplementary Fig. 15). Additionally, the interval between first *P. aeruginosa* positive culture and first mucoid *P. aeruginosa* culture was significantly shorter among 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 we excluded study subjects who underwent exome sequencing from the validation analysis, the p-values from the validation study are independent of the p-value estimated from the exome sequence data. These results support a strong association between presence of *DCTN4* missense variants and susceptibility to both earlier and more severe *P. aeruginosa* infection.

Dynactin 4 is a component of the dynein-dependent motor that moves autophagosomes along microtubules into lysosomes for degradation as part of the autophagy process–a highly conserved cellular quality control mechanism to transport and degrade damaged proteins and microbes[14,15]. *P. aeruginosa* induces autophagy in alveolar macrophages *in vitro* and autophagy plays an essential role in the clearance of *P. aeruginosa*[16]. In cystic fibrosis, intracellular accumulation of Δ*F508-CFTR* is associated with reduced macroautophagic flux via inhibition of autophagosome formation. This results in increased airway inflammation[15]. 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 (e.g. Δ*F508*) 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, neither would have achieved genome-wide significance via GWAS, and neither variant is on common SNP-genotyping platforms or tagged well (Supplementary Note). Our success with exome sequencing was due in part to the synergy of several important factors: phenotypically well-matched extremes with the exception of the trait of interest, the effect size estimated for *DCTN4* is relatively large, the collective MAF for implicated variants is reasonably high (0.065) and we intentionally included these higher MAF variants 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.

Online Methods

Exome sequencing

QC of sample DNA—Initial quality control (QC) performed on all samples included sample quantification (PicoGreen), confirmation of high-molecular weight DNA, test PCR amplification (four amplicons), and sex determination using a Taq-man assay[17]. 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 through sample preparation and the sequencing pipeline.

Library production and exome capture—Approximately 3.5 ug of genomic DNA was used for a series of shotgun library construction steps, including fragmentation through 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 qPCR protocol (Kapa Biosystems). Using the automated Illumina cBot cluster station, non-multiplexed samples were processed in batches of eight (one for each lane of the flow-cell), diluted and denatured to their final effective loading concentrations. Hybridization was followed by cluster generation via bridge PCR as per standard protocols (Illumina). Enriched libraries were sequenced on an Illumina GAIIx using paired-end 76 base runs.

Read mapping and variant analysis—Samples were processed from real-time base-calls (RTA 1.7 software [Bustard], converted to qseq.txt files, and aligned to a human reference (hg19) using BWA (Burrows-Wheeler Aligner)[18]. Read-pairs not mapping within ±2 standard deviations of the average library size (~125 ± 15 bp for exomes) were removed. Data were processed using the Genome Analysis ToolKit (GATK refv1.2905[19]). All aligned read data were subjected removal of reads with duplicate start positions, indel realignment and base qualities recalibration. Variant detection and genotyping were performed using the UnifiedGenotyper (UG) tool from GATK and only performed on the targeted exome regions. Variant data for each sample was formatted (variant call format [VCF]) as "raw" calls for all samples, and sites flagged to mark sites of lower quality/false positives (i.e. low quality scores (≤50), allelic imbalance (≥0.75), long homopolymer runs (>3), and/or low quality by depth (QD<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 QC—Individual exome sequencing data were evaluated against the QC metrics (Supplementary Table 1) including assessment of: (1) total reads: a minimum of 30M PE reads; (2) library complexity: the ratio of unique reads to total reads mapped to target; (3) capture efficiency: the ratio of reads mapped to target versus the reads mapped to human; (4) coverage distribution: 90% at ≥8× required for completion; (5) capture uniformity; (6) raw error rates; (7) Ti/Tv ratio (3.2 for known sites and 2.9 for novel sites); (8) distribution of known and novel variants relative to dbSNP; (9) fingerprint concordance >99%; (10) homozygosity; and (11) heterozygosity. All QC metrics for both single-lane and merged data were reviewed to identify data deviations from known or historical norms. Lanes/samples that failed QC were re-queued for library prep or further sequencing. Variants sites failing the following criteria were excluded from analyses: > 10%

missingness, failing Hardy Weinberg Equilibrium (p< 0.005/N), Qual < 30, QD < 5, AB < .25 or AB > 0.75 (Supplementary Table 1). No *DCTN4* sites for any individuals were removed due to low quality.

Statistical analysis

Selection of phenotypic extremes—The definition of chronic *Pseudomonas aeruginosa* (*Pa*) 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[8]. We defined an individual to have reached the chronic endpoint if s/he had a 6-month period of positive cultures. This definition also is very similar to the "Leeds" criterion suggested by Lee et al.[12] Individuals in the analysis set had a median of 3.5 culture-quarters per year.

Selection of Extreme Individuals—Of the 38 early onset chronic *P. aeruginosa* EPIC individuals in the ES sample, half of these early onset individuals were in the earliest 5% overall (≤2.5 years at onset) and all were in the earliest 7% (≤ 5 years, as estimated by the Kaplan Meier curve using the 1322 exome-eligible participants from EPIC DNA Collection study). These individuals represent the worst extreme in terms of overall frequency of *P. aeruginosa* positive quarters, all being in the worst 20% and half among the worst 4% (>24% positive). These 38 individuals had a total of 1165 quarters of observation time, with 334 of these positive for *P. aeruginosa* (28%). The five GMS individuals selected for the early onset *P. aeruginosa* extreme had continuous positive culture-years from the first or second year of life until the end of follow-up (minimum follow-up age = 12, maximum = 19) except for one year each for two individuals.

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. Never *P. aeruginosa* individuals were chosen to balance the early onset chronic sample on sex and *CFTR* clinical risk group (1, 2 or unknown), though most individuals were in risk group 1 (35 in the early onset extreme and 37 in the late extreme for successful exomes).[11] The 10 individuals selected for the late onset extreme from the 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—The method described by Morris and Zeggini[10] was used to obtain p-values for each gene. We adjusted each gene test for *CFTR* risk group (risk group 1 or not) and PC1, PC2 and PC18 from a principal components decomposition of the entire set of exome data after applying the QC filters described above (Supplementary Table 2). Indicators for sex and siblings were initially entered but discarded after the empirical distributions of covariates for these variables were found to center at zero. The first two PCs were included in the regression model to adjust for possible ancestral stratification. PC18 was added to the model after examining box plots of PCs 3 through 20 for cases versus controls, with PC18 showing a relatively large difference between cases and controls.

Resampling-based estimation of p-value—The nominal p-value for *DCTN4* from the MZ test, as well as the general performance of the test for these 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 based on the ESP-wide observed MAFs and correlation (rs11954652 MAF=0.048 and rs35772018 MAF =0.0178). Alleles were drawn for each person and locus, independently, under the null hypothesis of no association with phenotype group. The MZ 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 in order to obtain a precise estimate of the p-value as well as examine the behavior of the MZ 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 (HR) 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*, age at onset of first *Pa* positive culture, 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 (4 progressions to chronic *P. aeruginosa* in each), these last two groups were combined to gain more precision in the estimate of the effect size. All models were stratified on 5 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 in order to adjust for less sensitivity for detection of the end-point in individuals with fewer observations. Neither sex nor *CFTR*-Δ*508* homozygous genotype were significant predictors of any events (nor their interactions with the *DCTN4* variants score) and were not included in the final models. Race was not included in the model because there were not enough individuals on non-European ancestry to attain convergence of the estimation algorithm. However, the model was fitted for individuals of European ancestry only to determine whether the effects seen for *DCTN4* were driven by the few African American individuals in the analysis (Supplementary Note).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank the families for their participation and the EPIC study site investigators, research coordinators (Supplement Note) for their assistance. We thank A. Bigham, S. Leal, M. Rosenfeld, B. Ramsey, N. Hamblett, K. Buckingham, M. McMillin, S. McNamara, S. Ruuskafor technical assistance and helpful discussion. The authors wish to acknowledge the support of the National Heart, Lung, and Blood Institute (NHLBI) and the contributions of the research institutions, study investigators, field staff and study participants in creating this resource for biomedical research. Funding for GO ESP was provided by NHLBI grants RC2 HL-103010 (HeartGO), RC2 HL-102923 (LungGO) and RC2 HL-102924 (WHISP). The exome sequencing was performed through NHLBI grants RC2 HL-102925 (BroadGO) and RC2 HL-102926 (SeattleGO).Our work was supported in part by grants from the Cystic Fibrosis Foundation (to R.L.G [GIBSON07K0], and to Margaret Rosenfeld and RLG [CFF EPIC09K0]), the National Institutes of Health/National Human Genome Research Institute (5RO1HG004316 to

H.K.T.),and the Life Sciences Discovery Fund (2065508 and 0905001).K.C.B. was supported in part by the Mary Beryl Patch Turnbull Scholar Program.

Literature cited

1. Bamshad MJ, et al. Exome sequencing as a tool for Mendelian disease gene discovery. Nature reviews. Genetics. 2011; 12:745–755.
2. Stitziel NO, Kiezun A, Sunyaev S. Computational and statistical approaches to analyzing variants identified by exome sequencing. Genome biology. 2011; 12:227. [PubMed: 21920052]
3. Lanktree MB, Hegele RA, Schork NJ, Spence JD. Extremes of unexplained variation as a phenotype: an efficient approach for genome-wide association studies of cardiovascular disease. Circ Cardiovasc Genet. 2010; 3:215–221. [PubMed: 20407100]
4. Gibson RL, Burns JL, Ramsey BW. Pathophysiology and management of pulmonary infections in cystic fibrosis. American journal of respiratory and critical care medicine. 2003; 168:918–951. [PubMed: 14555458]
5. Emerson J, Rosenfeld M, McNamara S, Ramsey B, Gibson RL. Pseudomonas aeruginosa and other predictors of mortality and morbidity in young children with cystic fibrosis. Pediatric pulmonology. 2002; 34:91–100. [PubMed: 12112774]
6. Proesmans M, et al. Evaluating the "Leeds criteria" for Pseudomonas aeruginosa infection in a cystic fibrosis centre. The European respiratory journal : official journal of the European Society for Clinical Respiratory Physiology. 2006; 27:937–943.
7. Johansen HK, 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. Pediatric pulmonology. 2004; 37:427–432. [PubMed: 15095326]
8. Treggiari MM, et al. Early anti-pseudomonal acquisition in young patients with cystic fibrosis: rationale and design of the EPIC clinical trial and observational study'. Contemporary clinical trials. 2009; 30:256–268. [PubMed: 19470318]
9. Wright FA, et al. Genome-wide association and linkage identify modifier loci of lung disease severity in cystic fibrosis at 11p13 and 20q13.2. Nature genetics. 2011; 43:539–546. [PubMed: 21602797]
10. Morris AP, Zeggini E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet Epidemiol. 2010; 34:188–193. [PubMed: 19810025]
11. Green DM, et al. Mutations that permit residual CFTR function delay acquisition of multiple respiratory pathogens in CF patients. Respiratory research. 2010; 11:140. [PubMed: 20932301]
12. Lee TW, Brownlee KG, Conway SP, Denton M, Littlewood JM. Evaluation of a new definition for chronic Pseudomonas aeruginosa infection in cystic fibrosis patients. Journal of cystic fibrosis : official journal of the European Cystic Fibrosis Society. 2003; 2:29–34. [PubMed: 15463843]
13. Li Z, et al. Longitudinal development of mucoid Pseudomonas aeruginosa infection and lung disease progression in children with cystic fibrosis. JAMA : the journal of the American Medical Association. 2005; 293:581–588. [PubMed: 15687313]
14. Kimura S, Noda T, Yoshimori T. Dynein-dependent movement of autophagosomes mediates efficient encounters with lysosomes. Cell structure and function. 2008; 33:109–122. [PubMed: 18388399]
15. Haspel JA, Choi AM. Autophagy: A Core Cellular Process with Emerging Links to Pulmonary Disease. American journal of respiratory and critical care medicine. 2011
16. Yuan K, et al. Autophagy plays an essential role in the clearance of Pseudomonas aeruginosa by alveolar macrophages. Journal of cell science. 2012
17. Yu L, Martinez FD, Klimecki WT. Automated high-throughput sex-typing assay. BioTechniques. 2004; 37:662–664. [PubMed: 15517978]
18. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25:1754–1760. [PubMed: 19451168]
19. DePristo MA, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011; 43:491–498. [PubMed: 21478889]

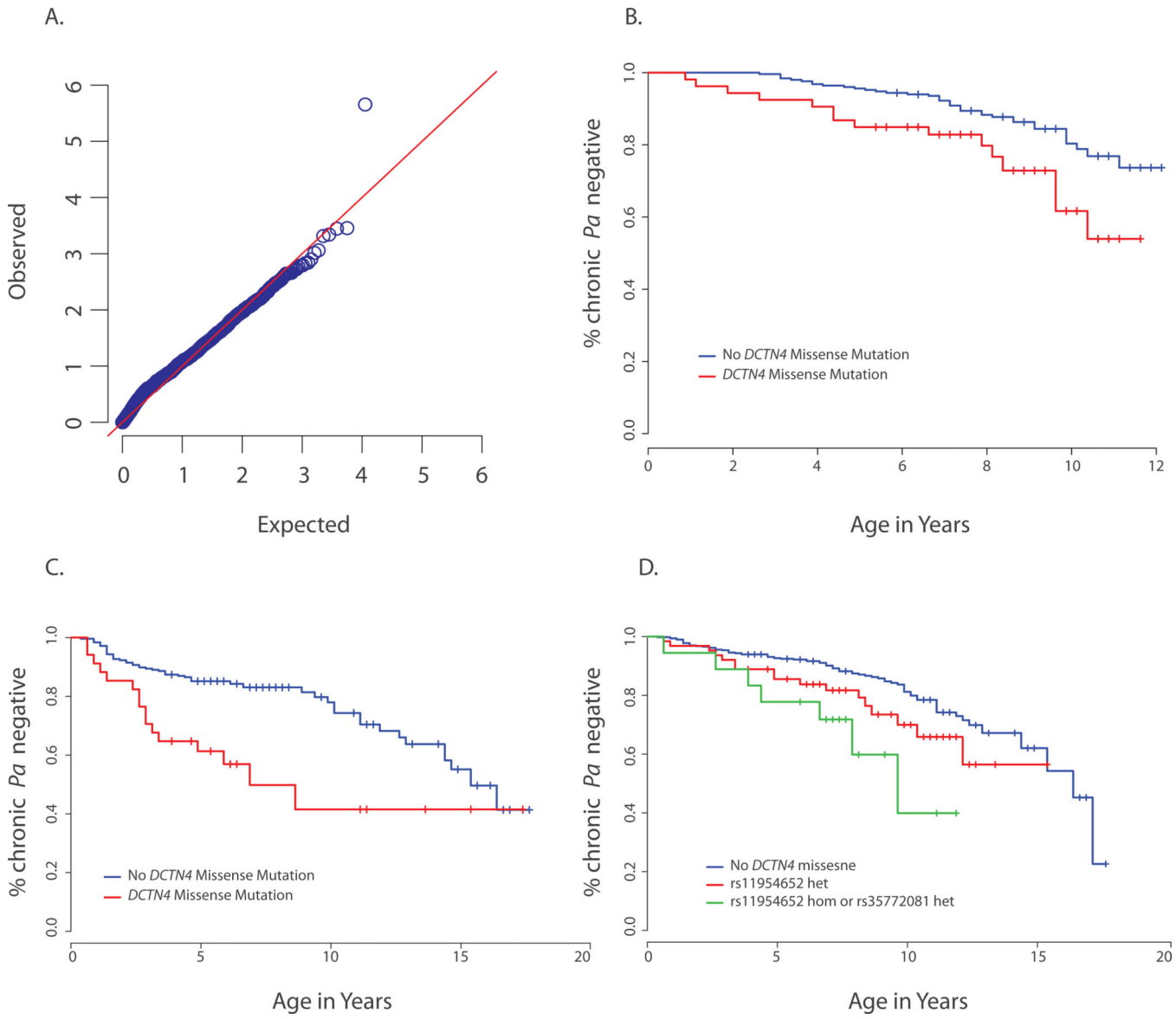

Fig. 1.

A) QQ-plot of p-values for a rare variant test of association (Online Methods) between each gene and extreme *P. aeruginosa* infection phenotypes. The most significant association ($p=2.2\times10^{-6)}$) is with *DCTN4* with a collapsed score based on two variants (rs11954652 and rs35772018). B) Kaplan-Meier curves comparing age-at-onset of chronic *P. aeruginosa* infection by presence of *DCTN4* variants among children in quintiles 2 and 3 of enrollment age among those reaching the endpoint (enrollment ages 1.6 to 6.7). Because of the need for analysis stratified on enrollment age, it is 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% 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 *P. aeruginosa* infection by presence of *DCTN4* variants among children who were not selected for a negative *P. aeruginosa*

history in the EPIC validation sample. Blue line: children without *DCTN4* variants (n=246); red dotted line: children with *DCTN4* variants rs11954652 and/or rs35772018 (n=34). HR=2.7 [1.4, 5.3] with p=0.004 (Online Methods). Comparison with 1B shows a larger baseline hazard for these children (both curves more steep than in 1B illustrates the need for stratification on enrollment age when implementing the Cox model. D) Kaplan-Meier curves comparing age-at-onset of chronic *P. aeruginosa* infection among all enrollment strata by *DCTN4* variant group; blue line: no *DCTN4* variants (n=565); red dotted line: rs11954652 heterozygotes (n=78); green line rs11954652 homozygotes and rs35772018 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 hazard ratio estimates because all enrollment strata are shown together in this plot: 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

Results of association analyses of *DCTN4* and *P. aeruginosa* phenotypes

Event	Group	N	# events	HR	95% CI	pvalue
Age at first Pa+ culture	no missense	565	345	1.0	--	--
	any missense	102	70	1.4	(1.1,1.8)	0.01
Age of onset of chronic Pa	no missense	565	93	1.0	--	--
	any missense	102	28	1.9	(1.2,2.9)	0.004
Age of onset of chronic Pa	No missense	565	93	1.0	(1.0, 2.7)	0.05
	Het rs11954652	78	20	1.7	(1.5, 6.9)	0.002
	Het rs35772018orhomrs 11954662	22	8	3.3		
Age of onset of mucoid Pa	no missense	565	84	1.0	--	--
	any missense	102	16	2.6	(1.1,5.9)	0.026
Time from first Pa to mucoid Pa	no missense	410	55	1.0	--	--
	any missense	89	17	3.8	(1.4, 10.5)	0.01