Additional Material

Fine-Scale Frequency of the *MUC5B* Promoter Variant Correlates with Idiopathic Pulmonary Fibrosis Healthcare Burden

Edmund Gilbert[1,2], Aoife Carolan[3,4], Mari Ozaki[3,4], Wan Lin Ng[3,5], Jisha Jasmin[3,6], David A Schwartz[7], Michael T Henry[8], Gianpiero L. Cavalleri[1,2], Killian Hurley[3,4], Irish Pulmonary Fibrosis Research Consortium*

Methods

Genotype Data

Using PLINK (1, 2)], we merged the Irish ancestry, genome-wide, SNP genotypes previously reported from the Irish DNA Atlas (3)[REF] and the Trinity Student (4) (dbGaP Study Accession: phs000789.v1.p1) datasets. New genotypes from the Irish DNA Atlas were added to this dataset, using the same genotyping procedure as previously described (3). Performing quality control of these genotypes, we removed individuals with a missingness >5%, then SNP with a missingness >5%, minor allele frequency <1%, and a p-value denoting deviation from Hardy-Weinberg expectations < 1e-9. To remove close relatedness from the dataset and thereby capture population-level relatedness and diversity, we further removed one from each detected close family pairs, using KING (5), prioritising participants from the Irish DNA Atlas due to their added geographic data. This left a final dataset of 2,604 individuals (Irish DNA Atlas n=381, Trinity Student n=2,223).

To cluster these 2,604 individuals, we phased autosomal genotypes using Beagle v5.4 (6) detected Identity-by-Descent segments with hap-ibd (7). For each pair of individuals, we summed the total length of IBD segments with an individual length of > 3 cM and < 30 cM, and then constructed a network object from these summed lengths in python and using the python implementation of the igraph package. We then performed community detection using the Leiden algorithm (8) using the python implementation in the leidenalg package, over two hierarchical levels of clustering. We detected the first "top" level of clusters initially, and then further sub-divided each of the four top level clusters into a second level of sub-clusters. Roughly, the top level corresponded to

Provincial differences, and the second level corresponded to County differences. We disregarded small second level clusters that were too small or contained too few Irish DNA Atlas individuals to annotate. This left a total of 2,465 individuals over 19 individual second level clusters, each grouped into one of 4 top level clusters.

The rs35705950 variant was not directly genotyped in either Irish reference dataset, therefore we sought to impute the genotype identity for each of our Irish Region Reference samples. Using the Michigan Imputation Server (9) and the Haplotype Reference Consortium panel (v2016), we imputed genotypes on chromosome 11. rs35705950 imputation was of high quality ($r^2 = 0.914$), matching the relatively high frequency of the variant overall. From these imputed genotypes, we estimated the allele frequencies in each second-level cluster using the PLINK functions --freq and --within. We then calculated confidence levels through a bootstrapping procedure. Briefly, for each bootstrap iteration, we randomly shuffled cluster assignments to Irish references – retaining original cluster sample size. Then we re-calculated the allele frequency for this pseudo-cluster. From the allele frequencies over 100 bootstrap replicates, we estimated the standard deviation and error for each cluster. Finally, to interpolate the allele frequencies of the rs35705950-T allele, we utilised Kriging in R using the Krig() function from the R package “fields”. As an input dataset, for each Irish DNA Atlas sample with geographic data, we assigned the rs35705950-T allele frequency based on the frequency estimated for the cluster that they were assigned to.

Discharge rates for IPF

We estimated the country wide health care use burden for IPF by extracting the discharge rates associated with the diagnosis of IPF (J841, ICD-10-AM/ACHI/ACS Eight

Edition https://www.ihacpa.gov.au/resources/icd-10-amachiacs-eighth-edition) using data from the Hospital Inpatient Enquiry (HIPE) supplied by The Healthcare Pricing Office, Ireland. To avoid changes in hospital admission and discharge patterns during the COVID-19 pandemic we examined hospital discharges from 2015 to 2019. We associated search of these discharges with the home county of the patient to estimate county-based burden. To account for population size differences in Irish regions, we also extracted county population sizes from the Irish 2016 census recorded by the Irish Central Statistics Office (CSO), selecting the population size of individuals self-identifying as "White Irish" to best proxy European-Irish ancestry. From these county-based discharge rates and population sizes we then estimated the proportion that each county contributes to the total number of discharges, and separately proportional population size. From these, we then estimated the county discharge rate weighted by population size by dividing the county's discharge proportion by the county's population size proportion. Then, to estimate each county's average allele frequency, we assigned each Irish DNA Atlas individual to the Irish county that their ancestral geographic position resides in, as well as recording the rs35705950-T allele frequency of that individual's cluster. Then for each county, we took the mean cluster rs35705950-T allele frequency using all Atlas participants assigned to that county by geographic position. Thereby controlling for counties with a mixture of genetic clusters, also accounting for uneven cluster representation in that county.

To control for the effects of age and sex on IPF risk when regressing county IPF burden with risk allele frequency, for each county in the Republic of Ireland we extracted the male-female ratio and average age as collected by the CSO in the 2016 census.

References:

1. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; 81: 559-575.

2. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 2015; 4: 7.

3. Gilbert E, O'Reilly S, Merrigan M, McGettigan D, Molloy AM, Brody LC, Bodmer W, Hutnik K, Ennis S, Lawson DJ, Wilson JF, Cavalleri GL. The Irish DNA Atlas: Revealing Fine-Scale Population Structure and History within Ireland. *Sci Rep* 2017; 7: 17199.

4. Desch KC, Ozel AB, Siemieniak D, Kalish Y, Shavit JA, Thornburg CD, Sharathkumar AA, McHugh CP, Laurie CC, Crenshaw A, Mirel DB, Kim Y, Cropp CD, Molloy AM, Kirke PN, Bailey-Wilson JE, Wilson AF, Mills JL, Scott JM, Brody LC, Li JZ, Ginsburg D. Linkage analysis identifies a locus for plasma von Willebrand factor undetected by genome-wide association. *Proc Natl Acad Sci U S A* 2013; 110: 588-593.

5. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. *Bioinformatics* 2010; 26: 2867-2873.

6. Browning BL, Tian X, Zhou Y, Browning SR. Fast two-stage phasing of large-scale sequence data. *Am J Hum Genet* 2021; 108: 1880-1890.

7. Zhou Y, Browning SR, Browning BL. A Fast and Simple Method for Detecting Identity-by-Descent Segments in Large-Scale Data. *Am J Hum Genet* 2020; 106: 426-437.

8. Traag VA, Waltman L, van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep* 2019; 9: 5233.

9. Das S, Forer L, Schonherr S, Sidore C, Locke AE, Kwong A, Vrieze SI, Chew EY, Levy S, McGue M, Schlessinger D, Stambolian D, Loh PR, Iacono WG, Swaroop A, Scott LJ, Cucca F, Kronenberg F, Boehnke M, Abecasis GR, Fuchsberger C. Next-generation genotype imputation service and methods. *Nat Genet* 2016; 48: 1284-1287.