(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
7 August 2014 (07.08.2014)

WIPO | PCT

(10) International Publication Number
WO 2014/121180 A1

(51) International Patent Classification:
C12Q 1/68 (2006.01) *C12N 15/11* (2006.01)

(21) International Application Number:
PCT/US2014/014395

(22) International Filing Date:
3 February 2014 (03.02.2014)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/759,820 1 February 2013 (01.02.2013) US

(71) Applicant: **THE UNIVERSITY OF CHICAGO** [US/US]; 6030 S. Ellis Avenue, Chicago, Illinois 60637 (US).

(72) Inventors: **NOTH, Imre**; 5016 S. Ellis Ave., Chicago, Illinois 60615 (US). **GARCIA, Joe**; 4425 Hacienda Del Sol Rd, Tucson, Arizona 85718 (US). **KAMINSKI, Naftali**; 227 Church St., Apt. 7H, New Haven, Connecticut 06510 (US).

(74) Agents: **FAHRLANDER, Jill A.** et al.; 150 S. Wacker Drive, Suite 620, Chicago, Illinois 60606 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM,

[Continued on next page]

(54) Title: GENETIC VARIANTS IN INTERSTITIAL LUNG DISEASE SUBJECTS

(57) Abstract: Disclosed are methods and kits for diagnosing or predicting risk for developing interstitial pulmonary fibrosis or predicting survival of individuals with interstitial pulmonary fibrosis.

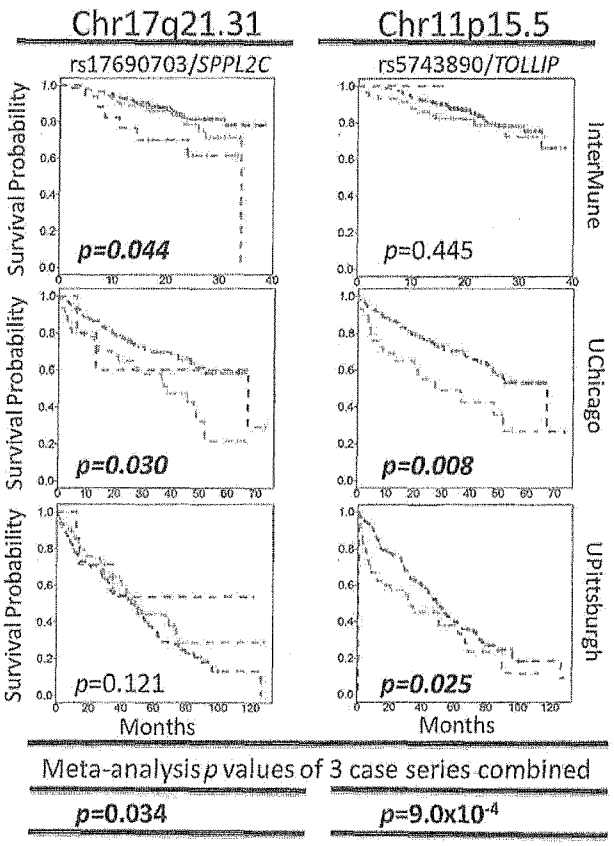

FIG. 1

WO 2014/121180 A1

DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

— *of inventorship (Rule 4.17(iv))*

Published:

— *with international search report (Art. 21(3))*

GENETIC VARIANTS IN INTERSTITIAL LUNG DISEASE SUBJECTS

STATEMENT REGARDING FEDERALLY SPONSORED RESEARCH OR DEVELOPMENT

Not applicable.

CROSS-REFERENCE TO RELATED APPLICATIONS

This PCT application claims the benefit of US Provisional Application No. 61/759,820, filed February 1, 2013, which is incorporated by reference herein.

INTRODUCTION

Idiopathic Pulmonary Fibrosis (IPF) is a low prevalence, devastating disease of unknown etiology characterized by an interstitial fibrotic process and high mortality. The course of disease is heterogeneous with a 2-5 year median survival from diagnosis. To date, lung transplantation remains the only successful treatment option, while immunosuppression regimens were recently demonstrated as harmful. Therefore, identifying genetic variants associated with susceptibility to IPF and alleles involved in the heterogeneity of disease course and mortality remains a major challenge.

A common single nucleotide polymorphism (SNP) of *MUC5B* is present in 34-38% of non-familial IPF cases, suggesting that a genetic underpinning contributes to disease. A prior genome-wide association study (GWAS) examining approximately 250,000 SNPs in 159 IPF cases demonstrated the association of an intronic common variant in telomerase reverse transcriptase (*TERT)* gene with susceptibility to IPF[1] Mutations in *TERT* or telomerase RNA component (*TERC*) genes result in telomere shortening and are associated with both familial and non-familial IPF. Rare heterozygous variants in surfactant protein A2 (*SFTPA2*) and surfactant protein C (*SFTPC*) genes have also been implicated in familial IPF. These findings suggest that the etiology of IPF may integrate multiple genetic loci.

There is a need in the art to identify genetic variants in interstitial lung disease subjects. Provided here are methods and compositions addressing these and other needs in the art.

SUMMARY OF THE INVENTION

In certain embodiments is provided compositions and methods for identifying genetic variants in interstitial lung disease subjects. Also provided are compositions and methods of determining whether a human subject has, or is at risk of developing, an interstitial lung disease. In certain embodiments, the methods include detecting whether the genome of the subject comprises a genetic variant of at least one of TOLLIP, SPPL2C, and MDGA2, the presence of the genetic variant indicating that the subject has or is at risk of developing the interstitial lung disease. In certain embodiments, more than one genetic variant of TOLLIP and/or SPPL2C and/or MDGA2 is detected. In certain embodiments, in addition to detecting genetic variants of TOLLIP and/or SPPL2C and/or MDGA2, the method includes detecting whether the genome of the subject includes other genetic variants diagnostic or predictive of risk for interstitial lung disease, e.g., a genetic variant of MUC5B, such as rs35705950.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 shows probability of survival over time, revealing the association with marker SNPs in the ch11p15.5 and ch17q21.31 regions on InterMune, UChicago and UPittsburgh case series. Brown=homozygote minor; green=heterozygote; blue=homozygote major for each single nucleotide polymorphism.

Fig. 2A is a flowchart showing the approach used in a three-stage association study;

Fig. 2B is flowchart of mortality analyses by regression.

Fig. 3. QQ plot of the genome-wide association study (GWAS) of idiopathic pulmonary fibrosis (IPF).

Fig. 4 includes regional association plots showing the IPF-associated regions in Ch11p15.5 (Fig. 4A) and Ch17q21.31 (Fig. 4B).

Fig. 5 survival probability over time for people with or without H2 and with or without an SPPL2C variant.

Fig. 6A is a KM plot for TOLLIP*/MUC5B risk alleles; Fig. 6B is KM plot by Risk Index for WPGS using all 3 genes (TOLLIP, SPPL2C & MUC5B) and categorizing into 4 groups.

Fig. 7A-7C is a list of top associated loci with susceptibility to IPF.

Fig. 8 is a table listing the sample sources and sizes used in a three stage study.

Fig. 9 shows the characteristics of IPF patients used in stage 1 discovery GWAS study.

Fig. 10 lists the characteristics of IPF patients by stage and availability.

Fig. 11A-11C is a list of 44 SNPs and their association p-values with susceptibility to IPF from stage 1, stage 2, and overall.

Fig. 12 shows characteristics of IPF case series for mortality analysis.

Fig. 13 is a table showing association signals with susceptibility to IPF across stages of six SNPs followed up in Stage 3.

Fig. 14 is a table listing SNP effects on mortality.

Fig. 15 provides summaries of univariate Cox analysis for mortality.

Fig. 16 provides summaries of univariate and multivariate Cox analysis for mortality

Fig. 17 provides summaries of Kaplan-Meier survival analysis.

Fig. 18 lists predictors of survival in IPF patients identified using a univariant Cox model.

Fig. 19A-Fig. 19B lists predictors of survival in IPF patients identified using a multivariate analysis of covariance.

Fig. 20 lists 30 regions identified showing the value of aggregation and using information in addition to protein coding SNPs, with the six p values represent highest-ranking SNPs in each region in bold.

DETAILED DESCRIPTION

As described in detail below, an independent genome wide association study (GWAS) was used to identify novel polymorphisms associated with IPF susceptibility and/or mortality. The association of two novel genetic loci and the replication of a third locus in a 3-stage association study are reported herein. These loci are also associated with mortality in case series with follow-up data.

Specifically, the results obtained identified three genetic loci and replicated the association of four novel SNPs (rs111521887, rs5743894, rs5743890, and rs17690703) in two novel loci (ch11p15.5/*TOLLIP* and ch17q21.31/*SPPL2C*), and the *MUC5B* promoter SNP (rs35705950) with IPF susceptibility in European-Americans through a three-stage case-control study. Another novel SNP (rs7144383) on a third genetic locus not previously known to be associated with IPF,

ch14q21.23/*MDGA2,* was discovered to show association with IPF susceptibility, although it did not replicate in Stage 3, possibly owing to the Stage 3 sample size.

The findings reported herein provide, *inter alia*, for novel compositions and methods for identifying genetic variants in interstitial lung disease subjects and/or determining whether an individual has, or is at risk for developing, interstitial lung disease and/or compositions and methods for predicting prognosis, e.g., survival time or mortality, of an individual with an interstitial lung disease, for example, a fibrotic interstitial lung disease, such as IPF, or familial interstitial pneumonia. Further, the identification of genetic loci and SNPs associated with interstitial lung disease contributes to the understanding of IPF pathogenesis and provides potential targets for novel treatment paradigms.

Definitions

Unless defined otherwise, technical and scientific terms used herein have the same meaning as commonly understood by a person of ordinary skill in the art. See, *e.g.*, Lackie, DICTIONARY OF CELL AND MOLECULAR BIOLOGY, Elsevier (4^{th} ed. 2007); Sambrook *et al.*, MOLECULAR CLONING, A LABORATORY MANUAL, Cold Springs Harbor Press (Cold Springs Harbor, NY 1989). The term "a" or "an" is intended to mean "one or more." The term "comprise" and variations thereof such as "comprises" and "comprising," when preceding the recitation of a step or an element, are intended to mean that the addition of further steps or elements is optional and not excluded. The following definitions are provided to facilitate understanding of certain terms used frequently herein and are not meant to limit the scope of the present disclosure.

The term "nucleic acid" refers to deoxyribonucleotides or ribonucleotides and polymers thereof in either single- or double-stranded form, and complements thereof. "Nucleic acid" or "oligonucleotide" or "polynucleotide" or grammatical equivalents used herein means at least two nucleotides covalently linked together. Oligonucleotides are typically from about 5, 6, 7, 8, 9, 10, 12, 15, 25, 30, 40, 50 or more nucleotides in length, up to about 100 nucleotides in length. Nucleic acids and polynucleotides are a polymers of any length, including longer lengths, *e.g.*, 200, 300, 500, 1000, 2000, 3000, 5000, 7000, 10,000, etc. The term "nucleotide" typically refers to a single unit of a polynucleotide, *i.e.*, a monomer. Nucleotides can be ribonucleotides, deoxyribonucleotides, or modified versions thereof.

As used herein, a "genetic variant" refers to a mutation, single nucleotide polymorphism (SNP), deletion variant, missense variant, insertion variant, inversion, or copy number variant.

The terms "probe" or "primer" refer to one or more nucleic acid fragments whose specific hybridization to a sample can be detected. A probe or primer can be of any length depending on the particular technique it will be used for. For example, PCR primers are generally between 10 and 40 nucleotides in length, while nucleic acid probes for, *e.g.*, a Southern blot, can be more than a hundred nucleotides in length. The probe or primers can be unlabeled or labeled as described below so that its binding to a target sequence can be detected (*e.g.*, with a FRET donor or acceptor label). The probe or primer can be designed based on one or more particular (preselected) portions of a chromosome, *e.g.,* one or more clones, an isolated whole chromosome or chromosome fragment, or a collection of polymerase chain reaction (PCR) amplification products. The length and complexity of the nucleic acid fixed onto the target element is not critical to the invention. One of skill can adjust these factors to provide optimum hybridization and signal production for a given hybridization and detection procedures, and to provide the required resolution among different genes or genomic locations.

Probes and primers can also be immobilized on a solid surface (*e.g.*, nitrocellulose, glass, quartz, fused silica slides), as in an array. Techniques for producing high density arrays can also be used for this purpose (*see*, *e.g.*, Fodor (1991) *Science* 767-773; Johnston (1998) *Curr. Biol.* 8: R171-R174; Schummer (1997) *Biotechniques* 23: 1087-1092; Kern (1997) *Biotechniques* 23: 120-124; U.S. Patent No. 5,143,854). One of skill will recognize that the precise sequence of particular probes and primers can be modified from the target sequence to a certain degree to produce probes that are "substantially identical" or "substantially complementary to" a target sequence, but retain the ability to specifically bind to (*i.e.*, hybridize specifically to) the same targets from which they were derived.

A probe or primer is "capable of detecting" a genetic variant if it is complementary to a region that covers or is adjacent to the genetic variant. For example, to detect a SNP, primers can be designed on either side of the SNP, and primer extension used to determine the identity of the nucleotide at the position of the SNP. In some embodiments, FRET-labeled primers are used (at least one

labeled with a FRET donor and at least one labeled with a FRET acceptor) so that FRET signal will be detected only upon hybridization of both primers. In some embodiments, a probe is used in conditions such that it hybridizes only to a genetic variant, or only to a dominant sequence. For example, the probe can be designed to hybridize to a junction point of a genetice inversion, but not to a sequence that does not include the inversion.

Again, in the context of nucleic acids, the term "capable of hybridizing to" refers to a polynucleotide sequence that forms non-covalent, Watson-Crick bonds with a complementary sequence. One of skill will understand that the percent complementarity need not be 100% for hybridization to occur, depending on the length of the polynucleotides, length of the complementary region, and stringency of the conditions. For example, a polynucleotide (*e.g.*, primer or probe) can be capable of hybrindizing (binding) to a polynucleotide having 80%, 85%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99% or 100% complementarity over the stretch of the complementary region. Stringency can be increased by reducing the length of the complementary region, reducing the G-C content of the complementary region, increasing temperature and/or detergent levels, varying salt levels and pH, etc. as known in the art. In some embodiments, a polynucleotide is capable of hybridizing to a complementary sequence in standard PCR annealing conditions. In the context of detecting genetic variants, the tolerated percent complementarity or number of mismatches will vary depending on the technique used for detection (see below).

In the context of nucleic acids, the term "amplification product" refers to a polynucleotide that results from an amplification reaction, *e.g.*, PCR and variations thereof, rtPCR, strand displacement reaction (SDR), ligase chain reaction (LCR), transcription mediated amplification (TMA), or Qbeta replication. A thermally stable polymerase, *e.g.*, Taq, can be used to avoid repeated addition of polymerase throughout amplification procedures that involve cyclic or extreme temperatures (*e.g.*, PCR and its variants).

The terms "label," "detectable moiety," "detectable agent," and like terms refer to a composition detectable by spectroscopic, photochemical, biochemical, immunochemical, chemical, or other physical means. For example, useful labels include fluorescent dyes, luminescent agents, radioisotopes (*e.g.*, ^{32}P, ^{3}H), electron-dense reagents, enzymes, biotin, digoxigenin, or haptens and proteins or other

entities which can be made detectable, *e.g.*, by affinity. Any method known in the art for conjugating a nucleic acid or other biomolecule to a label may be employed, *e.g.*, using methods described in Hermanson, Bioconjugate Techniques 1996, Academic Press, Inc., San Diego. The term "tag" can be used synonymously with the term "label," but generally refers to an affinity-based moiety, *e.g.*, a "His tag" for purification, or a "strepavidin tag" that interacts with biotin.

A "labeled" molecule (*e.g.*, nucleic acid, protein, or antibody) is one that is bound, either covalently, through a linker or a chemical bond, or noncovalently, through ionic, van der Waals, electrostatic, or hydrogen bonds to a label such that the presence of the molecule may be detected by detecting the presence of the label bound to the molecule.

Förster resonance energy transfer (abbreviated FRET), also known as fluorescence resonance energy transfer, is a mechanism describing energy transfer between two chromophores. A donor chromophore (FRET donor), initially in its electronic excited state, can transfer energy to an acceptor chromophore (FRET acceptor), which is typically less than 10 nm away, through nonradiative dipole-dipole coupling. The energy transferred to the FRET acceptor is detected as an emission of light (energy) when the FRET donor and acceptor are in proximity. A "FRET signal" is thus the signal that is generated by the emission of light from the acceptor. The efficiency of Förster resonance energy transfer between a donor and an acceptor dye separated by a distance of R is given by $E = 1/[1+(R/R_0)^6]$ with R_0 being the Förster radius of the donor-acceptor pair at which E= ½. R_0 is about 50-60 Å for some commonly used dye pairs (*e.g.*, Cy3-Cy5). FRET signal varies as the distance to the 6^{th} power. If the donor-acceptor pair is positioned around R_0, a small change in distance ranging from 1 Å to 50 Å can be measured with the greatest signal to noise. With current technology, 1 ms or faster parallel imaging of many single FRET pairs is achievable.

A "FRET pair" refers to a FRET donor and FRET acceptor pair that are capable of FRET detection.

The terms "fluorophore," "dye," "fluorescent molecule," "fluorescent dye," "FRET dye" and like terms are used synonymously herein unless otherwise indicated.

"Subject," "patient," "individual" and like terms are used interchangeably and refer to, except where indicated, humans and non-human animals. The term does not necessarily indicate that the subject has been diagnosed with a particular disease, but typically refers to an individual under medical supervision. A patient can be an individual that is seeking diagnosis, treatment, monitoring, adjustment or modification of an existing therapeutic regimen, etc.

As used herein, a "sample" refers to a biological sample obtained from a subject. Samples include material that is processed prior to carrying out testing, *e.g.*, genomic DNA separated or purified from other cellular and non-cellular debris. In the context of the present disclosure, the sample includes genomic DNA from the subject, *e.g.*, cheek swab, blood sample, mucosal sample, buccal swab, skin sample, hair, etc.

A "control" sample or value refers to a sample that serves as a reference, usually a known reference, for comparison to a test sample. For example, a test sample can be taken from a test condition, *e.g.*, a sample from an individual of unknown disease status, and compared to samples from individuals with known conditions, *e.g.*, healthy, or lacking a given genetic variation (negative control), or pulmonary disease or having a given genetic variation (positive control). A control can also represent an average value gathered from a number of tests or results. One of skill in the art will recognize that controls can be designed for assessment of any number of parameters. For example, a control can be devised to compare signal strength in given conditions, *e.g.*, in the presence of a test probe, or primer. One of skill in the art will understand which controls are valuable in a given situation and be able to analyze data based on comparisons to control values. Controls are also valuable for determining the significance of data. For example, if values for a given parameter are widely variant in controls, variation in test samples will not be considered as significant.

Diagnosis, prognosis, and treatment of interstitial lung disease

Provided herein are compositions and methods for determining whether a human subject has or is at risk of developing an interstitial lung disease and/or prognosing interstitial lung disease. In certain embodiments, the methods of the invention may be used in conjunction with any other diagnostic or prognostic criterion or method, including, but not limited to, currently known criterion or methods.

In certain embodiments, the method for determining whether a human subject has or is at risk of developing an interstitial lung disease includes detecting whether the genome of the subject comprises a genetic variant of at least one of TOLLIP, SPPL2C, and MDGA2, the presence of the genetic variant indicating that the subject has or is at risk of developing the interstitial lung disease. In certain embodiments, more than one genetic variant of TOLLIP and/or SPPL2C and/or MDGA2 is detected. In certain embodiments, in addition to detecting genetic variants of TOLLIP and/or SPPL2C and/or MDGA2, the method includes detecting whether the genome of the subject includes other genetic variants diagnostic or predictive of risk for interstitial lung disease, e.g., a genetic variant of MUC5B, such as rs35705950.

In some embodiments, the method for determining whether a human subject has or is at risk of developing an interstitial lung disease includes detecting the presence or absence of one or more SNPs selected from rs111521887, rs5743894, rs5743890, rs17690703, and rs7144383. The presence or absence of each SNP may be detected alone or in combination with each other, i.e., the methods of the invention may include detection of one, two, three, four, or five of rs111521887, rs5743894, rs5743890, rs17690703, and rs7144383 in any possible combination. In certain embodiments, the method includes detecting the presence or absence of from one to five of rs111521887, rs5743894, rs5743890, rs17690703, and rs7144383 in any combination and the presence or absence of any other SNP associated with an interstitial lung disease or its prognosis, including, without limitation, the *MUC5B* SNP rs35705950.

In some embodiments, the method for determining whether a human subject has or is at risk of developing an interstitial lung disease includes detecting the presence of rs111521887 (*e.g.*, G or other non-dominant allele). In some embodiments, the method for determining whether a human subject has or is at risk of developing an interstitial lung disease includes detecting the presence of rs5743894 (*e.g.*, G or other non-dominant allele). In some embodiments, the method for determining whether a human subject has or is at risk of developing an interstitial lung disease includes detecting the presence of rs5743890 (*e.g.*, G or other non-dominant allele). In some embodiments, the method for determining whether a human subject has or is at risk of developing an interstitial lung disease includes detecting the presence of rs17690703 (*e.g.*, T or other non-dominant allele).

In some embodiments, the method for determining whether a human subject has or is at risk of developing an interstitial lung disease includes detecting the presence of rs7144383 (*e.g.*, G or other non-dominant allele).

In certain embodiments, the method for determining whether a human subject has or is at risk of developing an interstitial lung disease includes detecting one or more genetic variants listed in Fig. 7. The one or more genetic variants may be detected alone or in any possible combination of from two to 52 of the listed genetic variants. If the method includes detecting rs35705950, then the method includes detecting at least one additional genetic variant from the remaining 51 genetic variants listed in Fig. 7.

In certain embodiments, the method includes prognosing an interstitial lung disease in a human subject. In certain embodiments, the method comprises detecting whether the genome of the subject comprises a genetic variant of TOLLIP and/or SPPL2C prognostic of increased or decreased survival. In certain embodiments, the methods include detecting whether the genome of the subject comprises a genetic variant of MUC5B and whether the genome comprises a genetic variant of a genetic variant of TOLLIP and/or SPPL2C prognostic of increased or decreased survival. In certain embodiments, the method includes detecting whether the genome comprises rs17690703 and/or rs5743890, each of which is predictive of decreased survival. In certain embodiments, the method detects whether the genome comprises rs35705950, which is predictive of increased survival, and rs17690703 and/or rs5743890. In some embodiments, the method comprises detecting rs17690703 (*e.g.*, T or other non-dominant allele), and prognosing reduced survival time for the subject, In some embodiments, the method comprises detecting rs5743890 (*e.g.*, G or other non-dominant allele), and prognosing reduced survival time for the subject.

In certain embodiments, the method for prognosing the interstitial lung disease in a human subject includes detecting one or more genetic variants listed in Fig. 7. The one or more genetic variants may be detected alone or in any possible combination of from two to 52 of the listed genetic variants. If the method includes detecting rs35705950, then the method includes detecting at least one additional genetic variant from the remaining 51 genetic variants listed in Fig. 7.

The present invention provides methods for detecting the presence or absence of at least one genetic variant in a human subject. In certain embodiments, the method includes detecting the presence or absence of at least one genetic variant of at least one of TOLLIP, SPPL2C, and MDGA2 in a sample from the subject. In certain embodiments, more than one genetic variant of TOLLIP and/or SPPL2C and/or MDGA2 is detected. In certain embodiments, in addition to detecting genetic variants of TOLLIP and/or SPPL2C and/or MDGA2, the method includes detecting a genetic variant of MUC5B, such as rs355950.

In certain embodiments, the method for detecting the presence or absence of at least one genetic variant in a human subject includes detecting the presence or absence of at least one genetic variant of the genetic variants listed in Fig. 7. The one or more genetic variants may be detected alone or in any possible combination of from two to 52 of the genetic variants listed in Fig. 7. If the method includes detecting rs35705950, then the method includes detecting at least one additional genetic variant from the remaining 51 genetic variants listed in Fig. 7. In certain embodiments, the at least one genetic variant includes one or more of a single nucleotide polymorphism selected from the group consisting of rs111521887, rs5743894, rs5743890, rs17690703, and rs7144383 in any possible combination.

In other embodiments, the method for detecting the presence or absence of at least one genetic variant in a human subject includes detecting the presence or absence of heterozygosity in least one genetic variant of the genetic variants listed in Fig. 7. Alternatively, the method for detecting the presence or absence of at least one genetic variant in a human subject includes detecting the presence or absence of homozygosity in least one genetic variant of the genetic variants listed in Fig. 7. The heterozygosity or homozygosity of the one or more genetic variants may be detected alone or in any possible combination of from two to 52 of the genetic variants listed in Fig. 7, wherein the genetic variant may be the same or different in the individual chromosomes present in the diploid human subject. If the method includes detecting heterozygosity or homozygosity of rs35705950, then the method includes detecting heterozygosity or homozygosity of at least one additional genetic variant from the remaining 51 genetic variants listed in Fig. 7. In certain embodiments, the heterozygosity or homozygosity of at least one genetic variant includes the heterozygosity or homozygosity of one or more of a single nucleotide

polymorphism selected from the group consisting of rs111521887, rs5743894, rs5743890, rs17690703, and rs7144383 in any possible combination.

Also provided is a method for testing for interstitial lung disease in a human subject that involves detecting the level of TOLLIP gene expression in a sample from the subject, a low level of TOLLIP gene expression relative to a control being indicative of interstitial lung disease. The level of gene expression may be detected by measuring, directly or indirectly, TOLLIP mRNA or by measuring Tollip protein by any suitable method, several of which are known in the art. The control may include, for example, a sample from a human that does not have interstitial lung disease or a value or set of values, for example, a normal range, derived from several humans that do not have interstitial lung disease. A low level of TOLLIP gene expression relative to a control (standard control) indicative of interstitial lung disease is a level that is less than about 50% of the control.

In certain embodiments, the present invention includes a method of treating a human subject having an interstitial lung disease comprising detecting the level of TOLLIP expression in a sample from the subject, and if the subject has a low level of TOLLIP expression relative to a control (standard control), administering to the subject an amount of a Tollip agonist, Tollip or a genetic construct expressing TOLLIP effective to treat the interstitial lung disease. An amount effective to treat the interstitial lung disease is an amount effective to delay onset, reduce frequency and/or severity of one or more symptoms, ameliorate one or more symptoms, and/or improve comfort and/or some function of the subject, e.g., respiratory function, relative to an untreated second subject or pool of subjects, or relative to, or to the same subject prior to treatment, or after cessation of treatment.

Methods of detecting a genetic variant

The methods of the invention are not limited to any particular way of detecting the presence or absence of a genetic variant (*e.g.* SNP) and can employ any suitable method to detect the presence or absence of a variant(s), of which numerous detection methods are known in the art.

Dynamic allele-specific hybridization (DASH) can be used to detect a genetic variant. DASH genotyping takes advantage of the differences in the melting temperature in DNA that results from the instability of mismatched base pairs. The process can be vastly automated and encompasses a few simple principles.

Typically, the target genomic segment is amplified and separated from non-target sequence, *e.g.*, through use of a biotinylated primer and chromatography. A probe that is specific for the particular allele is added to the amplification product. The probe can be designed to hybridize specifically to a variant sequence or to the dominant allelic sequence. The probe can be either labeled with or added in the presence of a molecule that fluoresces when bound to double-stranded DNA. The signal intensity is then measured as temperature is increased until the Tm can be determined. A non-matching sequence (either genetic variant or dominant allelic sequence, depending on probe design), will result in a lower than expected Tm.

DASH genotyping relies on a quantifiable change in Tm, and is thus capable of measuring many types of mutations, not just SNPs. Other benefits of DASH include its ability to work with label free probes and its simple design and performance conditions.

Molecular beacons can also be used to detect a genetic variant. This method makes use of a specifically engineered single-stranded oligonucleotide probe. The oligonucleotide is designed such that there are complementary regions at each end and a probe sequence located in between. This design allows the probe to take on a hairpin, or stem-loop, structure in its natural, isolated state. Attached to one end of the probe is a fluorophore and to the other end a fluorescence quencher. Because of the stem-loop structure of the probe, the fluorophore is in close proximity to the quencher, thus preventing the molecule from emitting any fluorescence. The molecule is also engineered such that only the probe sequence is complementary to the targeted genomic DNA sequence.

If the probe sequence of the molecular beacon encounters its target genomic DNA sequence during the assay, it will anneal and hybridize. Because of the length of the probe sequence, the hairpin segment of the probe will be denatured in favor of forming a longer, more stable probe-target hybrid. This conformational change permits the fluorophore and quencher to be free of their tight proximity due to the hairpin association, allowing the molecule to fluoresce.

If on the other hand, the probe sequence encounters a target sequence with as little as one non-complementary nucleotide, the molecular beacon will preferentially stay in its natural hairpin state and no fluorescence will be observed, as the fluorophore remains quenched. The unique design of these molecular beacons

allows for a simple diagnostic assay to identify SNPs at a given location. If a molecular beacon is designed to match a wild-type allele and another to match a mutant of the allele, the two can be used to identify the genotype of an individual. If only the first probe's fluorophore wavelength is detected during the assay then the individual is homozygous to the wild type. If only the second probe's wavelength is detected then the individual is homozygous to the mutant allele. Finally, if both wavelengths are detected, then both molecular beacons must be hybridizing to their complements and thus the individual must contain both alleles and be heterozygous.

A microarray can also be used to detect genetic variants. Hundreds of thousands of probes can be arrayed on a small chip, allowing for many genetic variants or SNPs to be interrogated simultaneously. Because SNP alleles only differ in one nucleotide and because it is difficult to achieve optimal hybridization conditions for all probes on the array, the target DNA has the potential to hybridize to mismatched probes. This can be addressed by using several redundant probes to interrogate each SNP. Probes can be designed to have the SNP site in several different locations as well as containing mismatches to the SNP allele. By comparing the differential amount of hybridization of the target DNA to each of these redundant probes, it is possible to determine specific homozygous and heterozygous alleles.

Restriction fragment length polymorphism (RFLP) can be used to detect genetic variants and SNPs. RFLP makes use of the many different restriction endonucleases and their high affinity to unique and specific restriction sites. By performing a digestion on a genomic sample and determining fragment lengths through a gel assay it is possible to ascertain whether or not the enzymes cut the expected restriction sites. A failure to cut the genomic sample results in an identifiably larger than expected fragment implying that there is a mutation at the point of the restriction site which is rendering it protected from nuclease activity.

PCR- and amplification-based methods can be used to detect genetic variants. For example, tetra-primer PCR employs two pairs of primers to amplify two alleles in one PCR reaction. The primers are designed such that the two primer pairs overlap at a SNP location but each matches perfectly to only one of the possible alleles. As a result, if a given allele is present in the PCR reaction, the primer pair specific to that allele will produce product but not the alternative allele with a different allelic sequence. The two primer pairs can be designed such that

their PCR products are of a significantly different length allowing for easily distinguishable bands by gel electrophoresis, or such that they are differently labeled.

Primer extension can also be used to detect genetic variants. Primer extension first involves the hybridization of a probe to the bases immediately upstream of the SNP nucleotide followed by a 'mini-sequencing' reaction, in which DNA polymerase extends the hybridized primer by adding a base that is complementary to the SNP nucleotide. The incorporated base that is detected determines the presence or absence of the SNP allele. Because primer extension is based on the highly accurate DNA polymerase enzyme, the method is generally very reliable. Primer extension is able to genotype most SNPs under very similar reaction conditions making it also highly flexible. The primer extension method is used in a number of assay formats, and can be detected using *e.g.*, fluorescent labels or mass spectrometry.

Primer extension can involve incorporation of either fluorescently labeled ddNTP or fluorescently labeled deoxynucleotides (dNTP). With ddNTPs, probes hybridize to the target DNA immediately upstream of SNP nucleotide, and a single, ddNTP complementary to the SNP allele is added to the 3' end of the probe (the missing 3'-hydroxyl in didioxynucleotide prevents further nucleotides from being added). Each ddNTP is labeled with a different fluorescent signal allowing for the detection of all four alleles in the same reaction. With dNTPs, allele-specific probes have 3' bases which are complementary to each of the SNP alleles being interrogated. If the target DNA contains an allele complementary to the 3' base of the probe, the target DNA will completely hybridize to the probe, allowing DNA polymerase to extend from the 3' end of the probe. This is detected by the incorporation of the fluorescently labeled dNTPs onto the end of the probe. If the target DNA does not contain an allele complementary to the probe's 3' base, the target DNA will produce a mismatch at the 3' end of the probe and DNA polymerase will not be able to extend from the 3' end of the probe.

The iPLEX® SNP genotyping method takes a slightly different approach, and relies on detection by mass spectrometer. Extension probes are designed in such a way that many different SNP assays can be amplified and analyzed in a PCR cocktail. The extension reaction uses ddNTPs as above, but the detection of the

SNP allele is dependent on the actual mass of the extension product and not on a fluorescent molecule. This method is for low to medium high throughput, and is not intended for whole genome scanning.

Primer extension methods are, however, amenable to high throughput analysis. Primer extension probes can be arrayed on slides allowing for many SNPs to be genotyped at once. Broadly referred to as arrayed primer extension (APEX), this technology has several benefits over methods based on differential hybridization of probes. Comparatively, APEX methods have greater discriminating power than methods using differential hybridization, as it is often impossible to obtain the optimal hybridization conditions for the thousands of probes on DNA microarrays (usually this is addressed by having highly redundant probes).

Oligonucleotide ligation assays can also be used to detect genetic variants. DNA ligase catalyzes the ligation of the 3' end of a DNA fragment to the 5' end of a directly adjacent DNA fragment. This mechanism can be used to interrogate a SNP by hybridizing two probes directly over the SNP polymorphic site, whereby ligation can occur if the probes are identical to the target DNA. For example, two probes can be designed; an allele-specific probe which hybridizes to the target DNA so that its 3' base is situated directly over the SNP nucleotide and a second probe that hybridizes the template upstream (downstream in the complementary strand) of the SNP polymorphic site providing a 5' end for the ligation reaction. If the allele-specific probe matches the target DNA, it will fully hybridize to the target DNA and ligation can occur. Ligation does not generally occur in the presence of a mismatched 3' base. Ligated or unligated products can be detected by gel electrophoresis, MALDI-TOF mass spectrometry or by capillary electrophoresis.

The 5'-nuclease activity of Taq DNA polymerase can be used for detecting genetic variants. The assay is performed concurrently with a PCR reaction and the results can be read in real-time. The assay requires forward and reverse PCR primers that will amplify a region that includes the SNP polymorphic site. Allele discrimination is achieved using FRET, and one or two allele-specific probes that hybridize to the SNP polymorphic site. The probes have a fluorophore linked to their 5' end and a quencher molecule linked to their 3' end. While the probe is intact, the quencher will remain in close proximity to the fluorophore, eliminating the fluorophore's signal . During the PCR amplification step, if the allele-specific probe is

perfectly complementary to the SNP allele, it will bind to the target DNA strand and then get degraded by 5'-nuclease activity of the Taq polymerase as it extends the DNA from the PCR primers. The degradation of the probe results in the separation of the fluorophore from the quencher molecule, generating a detectable signal. If the allele-specific probe is not perfectly complementary, it will have lower melting temperature and not bind as efficiently. This prevents the nuclease from acting on the probe.

Fluorescence resonance energy transfer (FRET) detection can be used for detection in primer extension and ligation reactions where the two labels are brought into close proximity to each other. It can also be used in the 5'-nuclease reaction, the molecular beacon reaction, and the invasive cleavage reactions where the neighboring donor/acceptor pair is separated by cleavage or disruption of the stem-loop structure that holds them together. FRET occurs when two conditions are met. First, the emission spectrum of the fluorescent donor dye must overlap with the excitation wavelength of the acceptor dye. Second, the two dyes must be in close proximity to each other because energy transfer drops off quickly with distance. The proximity requirement is what makes FRET a good detection method for a number of allelic discrimination mechanisms.

A variety of dyes can be used for FRET, and are known in the art. The most common ones are fluorescein, cyanine dyes (Cy3 to Cy7), rhodamine dyes (*e.g.* rhodamine 6G), the Alexa series of dyes (Alexa 405 to Alexa 730). Some of these dyes have been used in FRET networks (with multiple donors and acceptors). Optics for imaging all of these require detection from UV to near IR (*e.g.* Alex 405 to Cy7), and the Atto series of dyes (Atto-Tec GmbH). The Alexa series of dyes from Invitrogen cover the whole spectral range. They are very bright and photostable.

Example dye pairs for FRET labeling include Alexa-405/Alex-488, Alexa-488/Alexa-546, Alexa-532/Alexa-594, Alexa-594/Alexa-680, Alexa-594/Alexa-700, Alexa-700/Alexa-790, Cy3/Cy5, Cy3.5/Cy5.5, and Rhodamine-Green/Rhodamine-Red, etc. Fluorescent metal nanoparticles such as silver and gold nanoclusters can also be used (Richards *et al.* (2008) *J Am Chem Soc* 130:5038-39; Vosch *et al.* (2007) *Proc Natl Acad Sci USA* 104:12616-21; Petty and Dickson (2003) *J Am Chem Soc* 125:7780-81 Available filters, dichroics, multichroic mirrors and lasers can affect the choice of dye.

Kits

In certain embodiments, the present invention provides a kit for predicting, diagnosing, or prognosing interstitial lung disease in a human subject, the kit including (*e.g.* consisting essentially of) at least one probe or primer for detecting the presence or absence of at least one genetic variation. In certain embodiments, the at least one genetic variation includes a genetic variant of at least one of TOLLIP, SPPL2C, and MDGA2. In certain embodiments, the kit includes at least one primer or probe for detecting more than one genetic variant of TOLLIP and/or SPPL2C and/or MDGA2. In certain embodiments, the kit includes at least one probe or primer for detecting additional genetic variants diagnostic or predictive of risk for interstitial lung disease, *e.g.*, a genetic variant of MUC5B, such as rs37055950. In some embodiments, the kit includes a probe or primer for detecting one or more SNPs selected from rs111521887, rs5743894, rs5743890, rs17690703, and rs7144383. The kit may include probes or primers for detecting rs111521887, rs5743894, rs5743890, rs17690703, and rs7144383 alone or in any combination. In certain embodiments, the kit may include additional primers or probes for detecting the presence of detecting the presence or absence of rs37055950 and rs111521887, rs5743894, rs5743890, rs17690703, or rs7144383 in any combination. In certain embodiments, the kit includes at least one probe or primer includes at least one probe or primer for detecting one or more of the genetic variants listed in Fig. 7. The kit may include probes or primers for detecting the one or more genetic variants listed in Fig. 7 alone or in any possible combination of from two to 52 of the listed genetic variants. If the kit includes a probe or primer for detecting rs35705950, the kit also includes a probe or primer for detecting at least one additional genetic variant from the remaining 51 genetic variants listed in Fig. 7.

Claims directed to kits for predicting, diagnosing, or prognosing interstitial lung disease in a human subject "consisting essentially of" certain types of probes or primers is intended to capture kits that include probes or primers that are suitable primarily for detecting genetic variants associated with interstitial lung disease in humans, although the kits may also include additional probes or primers used as controls, for example, probes or primers for detecting housekeeping genes such β-actin, tubulin, or glyceraldehyde-3-phosphate dehydrogenase, for example. In this context, the use of the transitional phrase "consisting essentially of" is intended to

exclude arrays containing thousands of probes, the vast majority of which are unrelated to interstitial lung disease. In certain embodiments, the kits may include buffers, enzymes, labels, and the like, for example, for use in isolating DNA or mRNA, generating cDNA, or for amplifying and/or detecting and/or sequencing specific SNPs.

In some embodiments, the kit includes (or consists essentially of) a nucleic acid primer capable of hybridizing to a genetic variant in the TOLLIP gene (*e.g.*, a TOLLIP nucleic acid), SPPL2C gene (*e.g.*, a SPPL2C nucleic acid), or MDGA2 gene (*e.g.*, MDGA2 nucleic acid). In some embodiments, the genetic variant has been extracted from a human subject with an interstitial lung disease, or suspected of having an interstitial lung disease. In some embodiments, the genetic variant is an amplification product of DNA extracted from a human subject with an interstitial lung disease, or suspected of having an interstitial lung disease. In some embodiments, the interstitial lung disease is a pulmonary fibrotic condition.

In some embodiments, the kit includes a first nucleic acid probe (*e.g.*, a labeled probe) capable of hybridizing to an amplification product of a genetic variant in the TOLLIP gene (*e.g.*, a TOLLIP nucleic acid), SPPL2C gene (*e.g.*, a SPPL2C nucleic acid), or MDGA2 gene (*e.g.*, MDGA2 nucleic acid). In some embodiments, the kit includes a second nucleic acid probe capable of hybridizing to an amplification product of a genetic variant in the TOLLIP gene (*e.g.*, a TOLLIP nucleic acid), SPPL2C gene (*e.g.*, a SPPL2C nucleic acid), or MDGA2 gene (*e.g.*, MDGA2 nucleic acid). In some embodiments, the second nucleic acid probe is capable of hybridizing to a different sequence than the first probe. In some embodiments, only one of the nucleic acid probes hybridizes to the variant nucleotide(s) (*e.g.*, in the case of a SNP), while the other nucleic acid probe hybridizes to a nearby sequence. In some embodiments, the second probe is labeled, *e.g.*, with a different label than the first probe. In some embodiments, the first nucleic acid probe is labeled with a first label, and the second nucleic acid probe is labeled with a second label, wherein the first and second label form a FRET pair (are capable of fluorescence resonance energy transfer) when hybridized to the genetic variant TOLLIP gene (*e.g.*, a TOLLIP nucleic acid), SPPL2C gene (*e.g.*, a SPPL2C nucleic acid), or MDGA2 gene (*e.g.*, MDGA2 nucleic acid), or amplification product thereof.

In some embodiments, the kit includes (or consists essentially of) primers or at least one probe capable of detecting a genetic variant, *e.g.*, as described above, depending on the detection method selected. In some embodiments, the kit includes primers or at least one probe capable of detecting a genetic variant in a region selected from the group consisting of 11p15.5, 14q21.3, and 17q21.31. In some embodiments, the kit includes primers or at least one probe capable of detecting at least one genetic variant in 11p15.5 (*e.g.*, rs111521887, rs5743894, rs5743890, and rs35705950). In some embodiments, the kit includes primers or probes capable of detecting more than one (*e.g.,* 2, 3, 4, 5, 5-10, 10-20, or more) genetic variant in 11p15.5 and 14q21.3 (*e.g.*, rs7144383). In some embodiments, the kit includes primers or probes capable of detecting more than one (*e.g.,* 2, 3, 4, 5, 5-10, 10-20, or more) genetic variant in 11p15.5 and 17q21.31 (*e.g.*, rs17690703, a genetic inversion, or copy number variation). In some embodiments, the kit includes primers or probes capable of detecting more than one (*e.g.,* 2, 3, 4, 5, or more) genetic variant in 14q21.3 and 17q21.31. In some embodiments, the kit includes primers or probes capable of detecting more than one (*e.g.,* 2, 3, 4, 5, 5-10, 10-20, or more) genetic variant in 11p15.5, 14q21.3, and 17q21.31.

In some embodiments, the primers and/or probes are labeled, *e.g.*, with fluorescent labels or FRET labels. In some embodiments, the primers and/or probes are unlabeled. In some embodiments, the kit includes primers and/or probes that detect both a variant allelic sequence and the dominant allelic sequence at a selected genetic variant site, *e.g.*, with different labels, or designed to generate amplification or primer extension products with different masses.

In some embodiments, the kit further includes at least one control sample, *e.g.*, sample(s) with dominant allele(s) at the selected genetic variation site(s), or sample(s) with variant allele(s) at the selected genetic variation site(s). In some embodiments, the kit includes a polymerase.

***In vitro* complexes**

Provided herein are nucleic acid complexes, *e.g.*, formed in *in vitro* assays to indicate the presence of a genetic variant sequence. One of skill will understand that a nucleic acid complex can also be formed to detect the presence of a dominant allelic sequence, depending on the design of the probe or primer, *e.g.*, in assays to distinguish homozygous and heterozygous subjects.

In some embodiments, the complex comprises a first nucleic acid hybridized to a genetic variant nucleic acid, wherein the genetic variant nucleic acid is a genetic variant in a region selected from 11p15.5, 14q21.3, and 17q21.31. In some embodiments, the genetic variant nucleic acid is an amplification product. In some embodiments, the genetic variant nucleic acid is on genomic DNA, *e.g.*, from a subject that has or is suspected of having an interstitial lung disease. In some embodiments, the first nucleic acid is an amplification product or a primer extension product. In some embodiments, the first nucleic acid is labeled. In some embodiments, the nucleic acid complex further comprises a second nucleic acid hybridized to the genetic variant nucleic acid. In some embodiments, the second nucleic acid is labeled *e.g.*, with a FRET or other fluorescent label. In some embodiments, the first and second nucleic acids form a FRET pair when hybridized to a genetic variant sequence.

In some embodiments, the genetic variant is in the TOLLIP gene (*e.g.*, rs111521887, rs5743894, rs5743890). In some embodiments, the genetic variant is in the MDGA2 gene (*e.g.*, rs7144383). In some embodiments, the genetic variant is in the SPPL2C gene (*e.g.*, rs17690703, a genetic inversion, or copy number variation).

Further provided is an *in vitro* complex comprising a first nucleic acid probe (*e.g.*, a labeled probe) hybridized to a genetic variant nucleic acid, wherein said genetic variant nucleic acid comprises a genetic variant TOLLIP, SPPL2C or MDGA2 gene sequence, wherein said genetic variant nucleic acid is extracted from a human subject with an interstitial lung disease or suspected of having an interstitial lung disease, or is an amplification product thereof. In some embodiments, the complex further comprises a second nucleic acid probe (*e.g.*, labeled with a different label) hybridized to said genetic variant nucleic acid. In some embodiments, first nucleic acid probe comprises a first label and said second nucleic acid probe comprises a second label, wherein said first and second label are capable of fluorescence resonance energy transfer.

In some embodiments, the complex further comprises an enzyme, such as a DNA polymerase (*e.g.*, standard DNA polymerase or thermally stable polymerase such as Taq) or ligase.

Genetic variants associated with interstitial lung disease

MUC5B and *TOLLIP* genes reside on the same genetic locus. Based on the analysis performed, the association of *TOLLIP* genetic variants was found to be independent from association with the previously reported *MUC5B* promoter SNP, rs35705950, on IPF susceptibility. Notably, the minor allele of *TOLLIP* SNP, rs5743890_G, was discovered to be a "protective" allele, as it lowered susceptibility to IPF compared with controls. However, mortality analysis demonstrated that individuals who developed IPF despite having the protective rs5743890_G allele had increased mortality in two independent case series and in a meta-analysis. The *MUC5B*/*TOLLIP* region on chromosome 11p15.5 exemplifies the association patterns, disease susceptibility and outcomes.

The Toll interacting protein (Tollip), encoded by the *TOLLIP* gene, is known to be a critical regulator of Toll-like receptor (TLR)-mediated innate immune responses and transforming growth factor-β (TGF-β1) signaling pathway. Tollip activates Myd88-dependent NF-kB to modulate TLR signaling and membrane trafficking; interacts with Smad7 to modulate intracellular trafficking and negatively regulated TGF-β signaling pathway by degrading ubiquitinated TGF- β type 1 receptor; interacts with caveolin-1 interacting protein in monocytes, regulating signaling in antigen-presenting cells to induce antigen specific proliferation of T-cell proliferation, B cells, or both. *TOLLIP* polymorphisms are involved in regulation of TLR2 and TLR4 and are associated with susceptibility to tuberculosis, atopic dermatitis, sepsis, and *TOLLIP* is differentially hypomethylated in IPF lungs. Lastly, failure to upregulate *TOLLIP* expression in inflammatory bowel disease, may lead to chronic inflammation.

Chromosome 17q21 region has been associated with Parkinson's, multiple sclerosis, Alzheimer's, androgenic alopecia, and interestingly, with the response to inhaled corticosteroids in asthma and COPD. In the present study, it was discovered that the minor allele rs17690703_T in the 17q21.31 region was associated with decreased susceptibility for IPF development and also conferred increased mortality in InterMune, UChicago, and in the meta-analysis. Among the unique aspects of this region include a known inversion, referred to as H2, in a large region of conserved LD on the chromosome, which is positively selected in Europeans. There also appear to be a high number of copy number variants (CNVs) within this region and it

has been associated with a microdeletion syndrome. A critical span of 440 kb that partially or entirely involves five genes: *CRHR1*, *IMP5* (*SPPL2C*), *MAPT*, *STH* and *KIAA1267* reside on 17q21.31 region. A large number of variants in the region with significance in Stage 1 were discovered, with a focus on the top SNPs.

MDGA2, a novel region, resides on 14q21.23 and showed association with IPF susceptibility. *MDGA2* is a paralog for *ICAM,* which has been recently demonstrated as a potential biomarker of IPF disease activity. The instant findings indicate the importance of this gene in IPF.

IPF is a heterogeneous disease and, by definition, is a diagnosis of exclusion. As such, misdiagnoses are possible, which might lead to a reduction in power. However all subjects met currently accepted criteria for diagnosis as outlined by ATS/ERS/JRS/ELAT with many having been vetted with core pathology and radiology as in InterMune, ACE-IPF, as well as participation in variety of studies.

This discovery GWAS study revealed novel genetic loci associated with IPF susceptibility. Furthermore, susceptibility alleles within these loci were discovered to be associated with mortality. Identification of common genetic variants in association with IPF provides insight into the manifestations of this complex disease process and lead to earlier detection, more predictable prognosis, and personalized therapeutic strategies.

EXAMPLES

A three-stage association study was conducted including a discovery GWAS for susceptibility to IPF in Stage 1, and replicated the findings in two independent case-control association studies (Stage 2 and Stage 3, respectively). Association with mortality was evaluated in three case series. A flowchart illustrating the strategic approach used is shown in Fig. 2.

IPF cases and controls of each stage

Three stages of IPF cases were collected and characterized by the conventional criteria.[12-14] Stage 1 samples consisting of African-Americans (AA) and European-Americans (EA) were collected for the discovery phase of the genome-wide association study (GWAS), while Stages 2 and 3 consisting of only EA samples were collected for two independent replication studies (replication 1 and 2, respectively). All eligible subjects were at least 35 years of age and reported having symptoms of idiopathic interstitial pneumonia for at least 3 months. A high-resolution

computed tomographic scan was required to show definite or probable idiopathic interstitial pneumonia in accordance with predefined criteria,[14] and a surgical lung biopsy confirming UIP, was obtained in 37.3% of subjects in the discovery GWAS stage. Subjects with clinically significant exposure to known fibrogenic agents or another cause of interstitial lung disease were excluded.

Stage 1 discovery GWAS IPF samples (n=633) were identified and clinically characterized at the University of Chicago (UChicago), University of Pittsburgh (UPittsburgh), via the Lung Tissue Research Consortium (LTRC), and from the Correlating Outcomes with biomedical Markers to Estimate Time-progression in IPF (COMET) study. Stage 2 samples (n=544) comprised additional independent IPF patients from UChicago, InterMune,[3] Lung Transplant Outcomes Group (LTOG) cohort[4] and LTRC. Stage 3 IPF cases (n=324) consisted of additional independent IPF patients from LTOG and AntiCoagulant Effectiveness in Idiopathic Pulmonary Fibrosis Study (ACE-IPF).[5] Fig. 8, Fig. 9, and Fig. 10 feature each study population.

All eligible subjects were $\geq$ 35 years old and reported symptoms of idiopathic interstitial pneumonia for at least 3 months. A high-resolution computed tomographic scan was required for diagnosis of definite or probable idiopathic interstitial pneumonia in accordance with predefined criteria.[6] A surgical lung biopsy was obtained in 37.3% of affected subjects in the discovery GWAS stage. Subjects with clinically significant exposure to known fibrogenic agents and those with other known cause of interstitial lung disease were excluded.

For Stage 1, data of unaffected European American (EA) subjects, from dbGaP (n=1,442) were compiled with healthy subjects recruited from the University of Pittsburgh (n=103), to increase the available pool of subjects (n=1,545). A subset of controls matched one-on-one to cases by means of genome-wide genetic ancestry estimates were selected for downstream analysis.

EA controls for Stages 2 and 3 (n=687 and n=702, respectively) were collected from 2005 to 2012 as part of the Translational Research in the Department of Medicine Study (TRIDOM) at the University of Chicago. Institutional review boards at each institution approved this study and informed consent was obtained from all subjects. Summarized strategic methodology of the study and detailed clinical and demographic characteristics of all study stages are shown in Fig. 2 and Fig. 10, respectively.

Genotyping, imputation, and statistical analysis

Discovery Stage 1 genotyping was conducted using the Genome-Wide Human SNP 6.0 array (Affymetrix, Santa Clara, CA). Stages 2 and 3 genotyping was conducted using the iPLEX Gold™ Platform (Sequenom, San Diego, CA). Genotype imputation was performed with IMPUTE2 using European ancestry panel data from the 1000 Genomes Project as a reference. Association testing was performed using SNPTEST software (v2.3).[7] Fifty-two SNPs selected in 19 loci showing an association with IPF ($p<10^{-4}$) in Stage 1 were carried forward to Stage 2. As the selected SNPs with the lowest *p*-value in Stage 1 were all a result of imputation, their association was validated by genotyping using the iPLEX Gold™ Platform. Six SNPs in 3 loci achieving an overall $p<5x10^{-8}$ (i.e. Stage 1 and 2 combined) were carried forward to Stage 3.

DNA quantity was checked using PicoGreen fluorometry. Samples were dispensed at 50 ng/μl in 96-well plates and hybridized to arrays following manufacturer's protocols. Samples with fewer than 86% of the quality control (FQC) SNPs produced genotypes were rerun. Genotypes were recalled plate-by-plate in the study, including those downloaded from dbGaP using "crlmm" package, a new implementation of the Corrected Robust Linear Model with Maximum Likelihood Classification (CRLMM) algorithm, available through the Oligo package at Bioconductor.[18, 19]

Samples were excluded from the analysis if they failed any of several quality metrics: low call rate (below 97% or 93% for production plate with > 35 samples or with <35 samples, respectively), incompatibility between reported gender and genetically determined gender, or incompatibility between reported race and genetically determined race. Samples were also checked for unexpected familial relationships using pairwise IBD estimation in PLINK.[20] The total number of European-American IPF case and control samples passing all initial QC tests was 575 and 1,427 (1,340 of the available 1,442 cases from dbGaP and 87 of the 103 cases from University of Pittsburgh), respectively.

To reduce the false-positive rate, inflated by spuriously small *p*-values, while having little impact on the *p*-values associated with true positive loci for heterogeneous human populations, controls were matched to cases on a one-on-one basis for race and ethnicity based on genetic ancestry.[21] SMARTPCA

software,[22] was used to select control individuals from a larger set of available controls with the first four principal components (PCAs) obtained from a subset of variants showing limited linkage disequilibrium (n=267,000). To do so, the distance between every case individual and control individual was defined as the Euclidean distance between the individuals in a space based on the first four principal components, where each axis was also multiplied by its corresponding eigenvalue. After pairwise matching of 575 cases and 1,427 controls and accounted for the first four PCAs, 542 cases and 542 genetically matched controls were retained for downstream analysis.

Two tiers of filtering of control genotyping quality was performed using a call rate (<95%) and Hardy-Weinberg Equilibrium (HWE) p-value<10^{-3}. An additional 1,367 variants were further removed for inconsistent allele differences with IMPUTE2 1000 Genomes Project panel data. Prior to imputation, SNPs with minor allele frequency (MAF)< 5% were removed (a total of 349,801 were filtered based on QC and MAF) leaving a final number of 555,432 variants for further analysis and imputation.

Genome-wide SNP imputation was performed for the cleaned dataset to identify additional SNPs possibly showing associations. SHAPEIT[23] software was used to estimate phased haplotypes from the directly observed genotype data. Haplotypes derived from a European ancestry panel, consisting on samples from CEU, FIN, GBR, IBS and TSI from 1000 Genomes Project (February 2012 release), was used as a reference. Imputation was conducted using IMPUTE2. The inflation factor (λ) between cases and controls across all SNPs was 1.06. SNPTEST software (v2.3)[24] was used to calculate p-values based on a one degree-of-freedom score test for a logistic regression which assumes that the allele effect on the genotype for each SNP is additive. The score test implemented in SNPTEST allows for genotypic uncertainty via missing data likelihood, therefore it is applicable to both imputed genotypic data (i.e. in Stage 1) and to directly genotyped data (i.e. all stages). P-values were calculated for each stage separately, for Stages 1 and 2 combined, and finally for a joint analysis with all stages combined as one sample. Model parameters were estimated with a random subset of 200 individuals before imputation on the entire dataset.

Regions were deemed for follow-up in Stage 2 if they had a SNP with an association $p<10^{-4}$ in Stage 1. A minimum of 2 SNPs was selected from each region for Stage 2 genotyping. Where possible, the linkage disequilibrium (LD) of those two SNPs was low ($r^2<0.2$), where one of them was the variant with direct genotyping data showing the lowest *p*-value, and the other was the variant with imputed data showing the lowest *p*-value. Based on these criteria, a total of 40 SNPs for 19 loci were selected (2 SNPs per loci except for chr11/*TOLLIP*, chr17/*SPPL2C*, and chr7/*MAD1L1* regions with 3 SNPs; for chr7/*SHH* region with only 1 SNP).

In order to provide a better coverage of genetic variants for the previously reported region on chromosome 11p15.5, containing *TOLLIP* and *MUC5B*, an additional set of tagging SNPs (tSNPs) were selected using the multiple-marker selection algorithm, haplotype r^2, included in TagIT 3.03 software.[25] A set of 23 chr11/*TOLLIP* tSNPs under previously described criteria[26] from 380 European individuals (CEU, FIN, GBR, IBS and TSI) in 1000 Genomes Project Consortium[27] were selected. The common polymorphism of *MUC5B* (rs35705950) associated with familial and sporadic IPF cases was used as a positive control for genotyping quality and association. A total of 64 SNPs were compiled and submitted to Assay Design Suite (https://www.mysequenom.com/ToolsMassArray online design) for primers and probes design. Twelve of these SNPs failed during assay design and were considered failed and discarded from the analysis. A list of the remaining 52 SNPs from the 19 regions are shown in Fig. 7 along with their association *p*-values and MAFs.

A subset of 6 SNPs (rs111521887, rs17690703, rs35705950, rs5743890, rs5743894, rs7144383) from 3 loci showing a statistically significant association *p*-value$<5\times10^{-8}$ in the joint analysis of Stages 1 and 2 and with the same direction of effects in the two stages was compiled for Stage 3 replication (Fig. 11).

As the SNPs with the best *p*-value in the Stage 1 discovery GWAS were all a result of imputation, 541 of the 633 cases previously genotyped by the SNP array were compiled and genotyping was performed using iPLEX Gold™ platform to validate the findings. Approximately 10% of the samples were genotyped by TaqMan™ allelic discrimination assays (Applied Biosystems) to monitor genotyping quality. Genotyping was blind to case-control status. Samples with disconcordant genotypes were discarded.

Linkage disequilibrium assessment

Linkage disequilibrium (LD) between SNPs in the *MUC5B/TOLLIP* region was measured using pairwise r^2 measures.[8] The mode of inheritance for these SNPs (dominant, recessive) was determined by comparing the odds ratios of the heterozygous and at-risk homozygous genotypes. A regression-based conditional analysis of the interaction between *MUC5B* and *TOLLIP* SNPs on IPF susceptibility was implemented in the R statistical package.[9]

TOLLIP Gene expression in IPF lung tissues

Gene expression profiling data of IPF lungs was obtained from the Lung Genomics Research Consortium (LGRC) website. A total of 67 IPF individuals have paired genotype of SNPs associated with susceptibility and gene expression profiling data. The *TOLLIP* gene expression levels in these 67 samples were stratified into two groups according to presence or absence of the minor allele. Two-group comparison was performed using unequal variance *t*-test.

Mortality analysis for individual loci

Three case series in Stages 1 and 2 averaging follow-up data between 22 to 70 months (Fig. 12) were subjected to Cox regression analyses for mortality using the SPSS package (SPSS Inc., Chicago, IL) on the three IPF susceptibility loci that showed an overall $p<10^{-8}$ in Stages 1 and 2. Time "at risk" was defined as the interval between the date of enrollment in a given study and date of the last follow-up, lung transplant, or death. Lung transplant patients (2%, 7%, and 25% in InterMune, UChicago and UPittsburgh, respectively) were censored at time of transplant from the analysis, as potential confounders of survival. Univariate and multivariate analyses, considering relevant demographic and clinical parameters in the models, were conducted as appropriate. A single aggregate result for each locus was obtained by means of a meta-analysis applying both fixed and random effect models[10] as appropriate to account for the different available follow-up data among the case series studied.

Average follow-up data of 22 to 70 months was available for a subset of samples in 3 case series included in Stages 1 and 2 (Fig. 12). These case series were utilized mortality analyses was performed on the previously identified MUC5B promoter SNP and 5 novel SNPs within susceptibility loci that showed an overall association $p<10^{-8}$ in Stages 1 and 2 assuming that the genotypic effects were

additive. Logistic regressions were used initially to explore SNP effects comparing alive vs. dead patients. A more appropriate analysis of survival was then assessed on the 5 novel SNPs only, utilizing time "at risk".

All transplanted cases were censored from these analyses in order to avoid the confounding factor associated with IPF mortality. Univariate and multivariate analyses, using models considering relevant demographic and clinical parameters (such as age, gender, tobacco history, forced vital capacity (FVC) percent predicted, diffusing capacity of carbon monoxide (D_LCO) percent predicted, and recruitment center) were conducted. The heterogeneity of the Kaplan–Meier mortality curves as a function of genotypes for each SNP was assessed by the log-rank test. Hazard ratio (HR) estimates were obtained using Cox proportional hazard analyses. Schoenfeld residuals were used to assess the assumption of proportional hazards.

A single aggregate result for each locus was obtained with METASOFT by means of a meta-analysis. For that, both fixed and random effect models were applied, the latter corresponding to an optimized model to detect associations under heterogeneity, which was applied if heterogeneity between study samples was evident, as indicated by the significance of the Cochran's Q statistic.

Sample characteristics

Demographic and clinical characteristics of IPF patients and controls in each stage are shown in Fig. 9 and Fig. 10. As in other studies, cases in the discovery stage had a wide range of disease severity and age. The Stage 2 patients were a blend of cases with milder (InterMune) and more severe disease undergoing lung transplantation (LTOG), yielding a very similar group to Stage 1 based on the overall physiologic severity as assessed by forced vital capacity (FVC) and diffusing capacity for carbon monoxide (D_LCO) (Fig. 10). The Stage 3 patients were more severe, derived from the LTOG and ACE-IPF study. However all IPF cases met diagnostic criteria[16] and were all of similar age and gender. Characteristics of cases with follow-up data for survival analysis are shown in Fig. 12.

Genome-wide association study, replication, and regional association

After completion of sample quality control and genotype filtering, 542 of the 633 cases and 542 genetically matched controls selected from the available 1,545-pooled controls were retained for Stage 1. A total of 555,432 high quality genotyped variants were used for imputation which resulted in 10,601,812 best imputed

common variants with minor allele frequency (MAF) > 5%. The GWAS was then conducted using the genotyped and imputed SNPs. The inflation was modest with a test statistics of λ=1.06, indicating an insignificant confounding of the results by population stratification (Fig. 3).

A total of 19 genomic loci with an association (p-value<10^{-4}) were identified from Stage 1 discovery GWAS. Fifty-two SNPs were compiled from the combination of genotyped, imputed, and tSNPs. Fig. 7 summarizes annotations for these loci, allele frequency in reference populations (CEU, EUR), IPF cases, controls, as well as their association p-values with susceptibility to IPF.

Directly genotyped SNPs in Stage 2 nominally replicated many of the associations with IPF susceptibility detected in Stage 1 GWAS. Five imputed SNPs and the previously identified MUC5B promoter SNP reached genome-wide significance levels (p-value<4.2 x 10^{-8}) in a joint analysis of Stage 1 and 2. These six SNPs were re-genotyped in Stage 1 samples and the association confirmed. Fig. 11 highlighted loci of chr11p15.5 containing SNPs of *TOLLIP* (rs111521887, rs5743894, rs5743890) and *MUC5B* (rs35705950); chr17q21.31 of *SPPL2C* (rs17690703) and chr14q21.3 of *MDGA2* (rs7144383).

In Stage 3, the association of four of the SNPs in two novel loci (ch11p15.5/*TOLLIP* and ch17q21.31/*SPPL2C*) was replicated, as well as the association of *MUC5B* promoter SNP, previously reported in an independent study,[11] with IPF susceptibility. Each of them had overall combined p<10^{-9}, showing effects in the same direction across all single stages (i.e. allele rs35705950_T constitutes as risk for IPF, while alleles rs5743890_G and rs17690703_T protect from IPF) (Fig. 13). Regional associations of the genotyped and imputed SNPs at ch11p15.5 and ch17q21.31 loci are shown in Fig. 4. (A) ch11p15.5/*MUC5B*/*TOLLIP* locus and (B) ch17q21.31/*SPPL2C locus* as defined by the positions of SNPs showing a linkage disequilibrium with the lead SNP rs5743894 (A; p=2.2 × 10^{-6}) and SNP rs17690703 (B; p=4.9 × 10^{-6}), respectively. Disease associations as indicated by −log10 p-values are plotted against chromosomal positions. Diamonds and circles represent individual SNP of the GWA screen using genotyped and imputed data, respectively. Colored diamonds indicate SNP data obtained by the analysis of 542 IPF cases and 542 controls. Additional tSNPs selected for better coverage are included. Associations were assessed assuming recessive and additive modes of inheritance

for the *MUC5B/TOLLIP* locus and the *SPPL2C* locus, respectively. Levels of linkage disequilibria (r^2) with the best-associated SNP (red diamonds) are color-coded. Blue lines indicate recombination fractions as estimated from the European panel sample. Horizontal arrows mark structural human genes as annotated by Human Genome Build 37.3/gh19 of the UCSC (Genome Bioinformatics Group, University of California, Santa Cruz). Symbols, position and direction of each gene within the loci are shown at the bottom of the plot.

In ch11p15.5 locus, the r^2 values of *MUC5B* promoter SNP, rs35705950, and *TOLLIP* SNPs (rs111521887, rs5743894, and rs5743890) were 0.07, 0.16, and 0.01, respectively. These low levels of LD indicate that the signals of association for *TOLLIP* SNPs are independent from *MUC5B* (Fig. 4A). Moreover, the mode of effect for the *MUC5B* SNP (dominant) was different than that for the *TOLLIP* SNPs (additive or recessive), providing additional evidence that these are independent signals. Lastly, in a conditional regression-based analysis, genotypes were combined according to the mode of inheritance and it was found that, while the *MUC5B*/rs35705950 SNP showed the strongest signal ($p=2x10^{-16}$), the *TOLLIP*/rs11152887/rs5743894/rs5743890 SNPs remained associated with IPF ($p=0.05$).

Relationship between presence of susceptibility alleles by genotype and survival in IPF case series

Enrollment criteria in the InterMune study skewed patients towards better pulmonary function as assessed by FVC (71.56±12.68 percent predicted), and less heterogeneity of disease severity as assessed by a lesser standard deviation on lung function than in the UChicago study (65.17±18.29) or the UPittsburgh study (65.27±19.72). Also, InterMune had a shorter average follow-up period (22 months) in survivors, than UChicago (40 months) or UPittsburgh (70 months) (Fig. 12). Since the follow-up time varied widely in each IPF case series, it was decided to evaluate the novel susceptibility alleles in association with mortality, both separately and jointly through a meta-analysis.

Three SNPs were associated with mortality in an initial logistic regressions analysis of the overall case series (Fig. 14). Univariate Cox regression analysis in InterMune, UChicago, UPittsburgh as well as in the meta-analysis further demonstrated that the novel risk alleles for susceptibility in 11p15.5/*TOLLIP* and 17q21.31/*SPPL2C* loci were associated with protection from IPF mortality (Fig. 1 and

Fig. 15). Briefly, allele rs5743890_G was associated with increased mortality in UChicago (p=0.008) and in UPittsburgh (p=0.025). Similarly, allele rs17690703_T was associated with increased mortality in InterMune (p=0.044) and in UChicago (p=0.030). Meta-analysis of the 3 case series sustained associations with mortality (p=0.034 for rs17690703_T, and p=0.0009 for rs5743890_G) (Fig. 15). Notably, the meta-analysis of rs17690703_T with increased mortality suggested significant study heterogeneity among the three case series (Cochran's Q-value=9.54, p=0.0085). Multivariate analyses adjusting for recorded covariates (i.e. age, gender, tobacco history, FVC, D_LCO, at each recruitment center) that maintained p-value<0.1 in regression models did not appreciably change these findings (Fig. 16). Results of additional analyses pertaining to survival are presented in Fig. 17, Fig. 18, and Fig. 19.

The *SPPL2C* variant rs17690703 failed to meet significance only after adjustment of disease severity (p=0.06). This is highly suggestive that the region may have a relationship to survival. Intronic variants are unlikely to be causal. In fact, this variant might actually be a tag for an altogether different gene within the H2 inversion. Because H2 is rare among individuals of African (6%) and Asian (1%) ancestry, and the IPF cohort was overwhelmingly comprised of individuals of European ancestry (EA), where H2 occurs in approximately 20%, further evaluation of the role of H2 focused on an EA group. H2-specific SNPs tag the inversion, and are strongly correlated, but incompletely linked, to the *SPPL2C* variant (rs17690703) (r2 = 0.76). Three SNPs that tag H2 (rs916793, rs2902662, rs17651213) were included on the Affymetrix 6.0 GeneChip® (Affymetrix, Santa Clara, CA) used in the GWAS. Several proxy SNPs in complete linkage disequilibrium (LD) (r2 = 1) with SNPs that tag H2 were also identified. Included in this analysis were 120 EA individuals from the University of Chicago cohort for which mortality and genotype data were available. Of this group, 28.3% (n=34) carried an H2 haplotype, a 40% increase over the general population estimate. Of these 34 patients, 30 (88%) were heterozygous and 4(12%) homozygous for the inversion. Assignment of H2 status was based on the presence of all 3 SNPs that tag H2 (rs916793, rs2902662, rs17651213). This method allowed H2 assignment to all but 3 patients in this cohort. The addition of a proxy SNP (rs199448) allowed H2 status to be determined for the 3

remaining patients. These data suggest that presence of an H2 haplotype increases susceptibility to IPF.

To perform the survival analysis, the cohort of 120 EA individuals was then stratified based on H2 (absent vs. present) and *SPPL2C* (wild-type (WT) vs variant (Var)) status. Inclusion of *SPPL2C* in the stratification is necessary given the strong correlation between the two variants and potential confounding by *SPPL2C*. Survival analysis for each group showed a statistically significant difference (p=0.04) in mortality risk between the 4 groups (Fig. 5). The vast majority of patients belonged to either the H2(-)/*SPPL2C*-WT or H2(+)/*SPPL2C*-Var group, making it difficult to draw a conclusion about the two smaller groups. When comparing one group to another the statistical significance was lost. But these data suggest that H2 and *SPPL2C* contribute to mortality risk independently, *SPPL2C* (wild-type (WT) vs variant (Var)) status. Inclusion of *SPPL2C* in the stratification is necessary given the strong correlation between the two variants and potential confounding by *SPPL2C*.

The presence of a common variant associated with IPF, its location within an inversion, the independence of chromosome 17 from chromosome 11 and the possibility of variants related to survival in IPF indicate that additional sequencing may facilitate identification of causal or regulatory variants within the region.

A barrier inherent in the large amount of data generated by next generation sequencing of genetic regions involves methods to evaluate uncommon or rare variants. However the importance of regions and likelihood of additional uncommon and rare variants can be discovered by using aggregating or collapsing methods within regions. Indeed, regions with common variants have a greater number of uncommon or rare variants as well. One approach using the fundamentals of a logistic regression involves an L1-regularized regression to accommodate large number of variants. The Lasso method is a shrinkage and selection method for linear regression. It minimizes the usual sum of squared errors, with a bound on the sum of the absolute values of the coefficients. It has connections to soft-thresholding of wavelet coefficients, forward stagewise regression, and boosting methods. It is recognized that the power to detect true rare variants is limited by the number of cases (n=542) conducted on the array. A preliminary analysis based on aggregating the common and uncommon variants by region, ranking them by p values at 10^{-3} or smaller to yield expected rates of variants beyond that of the populace in general

was therefore conducted. The 542 cases versus controls were examined using a mutli-variant genetic association test of functional features of the genome by LASSO. This represents 37,000 functional features, which include but are not limited to protein coding regions as well as known lincRNA and miRNA, etc. The genome was analyzed in 5 MB increments or "clusters" where, 4.6 represents chromosome 4 and the 6 would then be cluster 35-30MB along that chromosome. Using $5x10^{-3}$ as a cutoff for significance, there were over 100 regions of interest. In Table X are the top 6 regions. Not surprisingly, MUC5B and TOLLIP in the same cluster were ranked second. Chromosome 17 actually demonstrated multiple clusters of interest. The preliminary data demonstrate several additional loci not previously identified, while also emphasizing the importance of 17q21.31 and 11p15.5 region, as well as other regions. This analysis demonstrates the ability to handle complex datasets of uncommon and rare variants to generate novel discovery.

To address the issue that variants within a region could exert an effect in opposite directions, a unique dataset with survival cohorts was used that that allows performance of linear regression analysis of each individual variant within the region to assess the direction of effect and assignment of an additive or subtractive model for multiple variants within a region. In fact, TOLLIP SNPs demonstrated this phenomenon, in that rs111521887 (G for T), or rs5743894 (G for T) of TOLLIP SNPs were associated with increased susceptibility to IPF while rs5743890 (G for T) was associated with decreased susceptibility, or a protective effect in developing IPF. However, all three SNPs seem to exert an effect reducing gene expression levels of TOLLIP in lung tissue, again arguing for a better understanding of causal or regulatory SNPs.

To further the integration of multiple genetic markers with clinical parameters, the two published SNPs in 11p15.5 were examined to determine if there was an interaction. The intersection of these two independent SNPs in TOLLIP and MUC5B demonstrated only a weak interaction with an r^2 of 0.009 by linear regression. The relationship with survival appears to therefore be additive in preliminary data. (Fig. 6A). Initial results indicated that the association with mortality for SPPL2C moved to only trend levels (p=0.06) after adjusting for severity of illness, with a modest hazard ratio of 1.3. However, taking into account information regarding the H1/H2 status and its influence, it is more than plausible that other SNPs in the regions will carry greater

hazard ratio of significance. Therefore the SPPL2C was incorporated into a risk index using a multidimensional approach to collapse categories down to 4 groups (Fig. 6B). An analysis using a weighted sum of risk index alleles across the SNPs where the Weighted Personalized Gene Risk Score (WPGRS) is obtained by multiplying the logHR by the number of risk alleles by genotype across 3 SNPs gives 17 categories. The unadjusted Cox regression model gave HR=6.51 (2.91-14.55), $p=5.02x10^{-06}$ and the adjusted Cox regression gave HR=6.60 (2.71-16.09), $p=3.23x10^{-05}$ (adjusted for age, sex, FVC, DLCO, study center) demonstrating the power of this approach. The identification of causal SNPs in the TOLLIP or SPPL2C region is expected to increase HR.

Novel genetic loci associated with IPF susceptibility.

In the joint analysis, two loci (ch11p15.5 and ch17q21.31) showed clear evidence of replication with effects in the same direction as in Stage 1 discovery GWAS and genome-wide significance levels of $p<10^{-8}$. Association of the genotyped and imputed SNPs at ch11p15.5 and ch17q21.31 loci is shown in **Figures S2A** and **S2B**, respectively. SNP rs35705950 on locus chr11p15.5/*MUC5B* has been firmly implicated in association with IPF.[17] Notably, three novel SNPs were revealed on the same locus, located in the intronic regions of *TOLLIP* gene, which were associated with IPF (rs111521887_G, Odds ratio (OR)=1.48, 95%CI=1.32-1.66, $p=2.2x10^{-12}$; rs5743894_G, OR=1.49, 95%CI=1.33-1.68, $p=1.35x10^{-12}$; rs5743890_G, OR=0.61, 95%CI=0.52-0.71, $p=3.43x10^{-11}$) (Fig. 13). Subsequent logistic regression analyses conditioned on the marker SNPs in ch11p15.5 revealed that these *TOLLIP* SNPs were not in LD with the *MUC5B* SNP, rs35705950. The r^2 values were 0.07, 0.16, 0.01 between rs35705950 and rs111521887, rs5743894, and rs5743890, respectively. This data indicated that the signals of association for these three SNPs were not correlated to rs35705950. Additionally, the mode of inheritance for the *MUC5B* SNP (dominant) is different than that for the *TOLLIP* SNPs (additive or recessive), adding to the evidence for independent signals. Lastly, genotypes were combined according to the mode of inheritance to identify the underlying genetic mode and perform a joint conditional analysis of rs35705950 and rs111521887: the *MUC5B* SNP shows the strongest signal ($p<2x10^{-16}$), but the *TOLLIP* SNP remains associated ($p=0.05$).

The second novel locus, which is located on chromosome 17q21.31, was indicated by imputation and supported by physical genotyping of SNP rs17690703_T (OR=0.70, 95%CI=0.62-0.79, p=5.70x10^{-9})(Fig. 13).

For the third novel locus on chromosome 14q21.3, replication of SNP rs7144383 was achieved in an independent case-control association study after imputation of the 1000 Genomes Project data demonstrating an OR=1.57, 95%CI=1.18-2.08, p=3.50x10^{-8} in the joint analysis. In a joint analysis of Stage 3 data along with data from the two previous stages this association maintained a suggestive association (OR=1.44, 95%CI=1.23-1.69, p=3.7x10^{-6}) (Fig. 13).

The following embodiments are included:

1. A method of determining whether a human subject has or is at risk of developing an interstitial lung disease, the method comprising detecting whether the genome of the subject comprises a genetic variant of at least one of TOLLIP, SPPL2C, and MDGA2 and determining whether the subject has or is at risk of developing an interstitial lung disease, the presence of the genetic variant indicating that the subject has or is at risk of developing the interstitial lung disease.

2. The method of embodiment 1, wherein the method comprises detecting whether the genome of the subject comprises a genetic variant of TOLLIP.

3. The method of embodiment 1 or embodiment 2, wherein the method comprises detecting whether the genome of the subject comprises a genetic variant of SPPL2C.

4. The method of any one of embodiments 1-3, wherein the method comprises detecting whether the genome of the subject comprises a genetic variant of MDGA2.

5. The method of any one of embodiments 1-4, further comprising detecting whether the genome of the subject comprises a genetic variant of MUC5B.

6. The method of any one of embodiments 1-5, wherein the method comprises detecting whether the genome of the subject comprises one or more genetic variants having a single nucleotide polymorphism selected from the group consisting of rs111521887, rs5743894, rs5743890, rs17690703, and rs7144383.

7. The method of embodiment 6, wherein the method comprises detecting whether the genome of the subject comprises the genetic variant having the single nucleotide polymorphism rs111521887.

8. The method of embodiment 6 or 7, wherein the method comprises detecting whether the genome of the subject comprises the genetic variant having the single nucleotide polymorphism rs5743894.

9. The method of any one of embodiments 6-8, wherein the method comprises detecting whether the genome of the subject comprises the genetic variant having the single nucleotide polymorphism rs5743890.

10. The method of any one of embodiments 6-9, wherein the method comprises detecting whether the genome of the subject comprises the genetic variant having the single nucleotide polymorphism rs17690703.

11. The method of any one of embodiments 6-10, wherein the method comprises detecting whether the genome of the subject comprises the genetic variant having the single nucleotide polymorphism rs7144383.

12. The method of any one of embodiments 6-11, further comprising detecting whether the genome of the subject comprises a genetic variant having a single polynucleotide polymorphism rs35705950.

13. The method of any one of embodiments 1-12, wherein the interstitial lung disease is a fibrotic interstitial lung disease.

14. The method of embodiment 13, wherein the interstitial lung disease is idiopathic pulmonary fibrosis or familial interstitial pneumonia.

15. A method of prognosing an interstitial lung disease in a human subject, the method comprising detecting whether the genome of the subject comprises a genetic variant of TOLLIP or SPPL2C and determining a prognosis for the subject, the presence of the genetic variant gene being prognostic of increased or decreased survival.

16. The method of embodiment 15, wherein the method comprises detecting whether the genome of the subject comprises a genetic variant of TOLLIP.

17. The method of embodiment 15 or 16, wherein the method comprises detecting whether the genome of the subject comprises a genetic variant of SPPL2C.

18. The method of any one of embodiments 15-17, further comprising detecting whether the genome of the subject comprises a genetic variant of MUC5B.

19. The method of any one of embodiments 15-18, wherein the genetic variant has at least one single nucleotide polymorphism selected from the group

consisting of rs17690703 and rs5743890, and wherein the single nucleotide polymorphism is predictive of decreased survival.

20. The method of any one of embodiments 15-19, wherein the genome of the subject comprises the single nucleotide polymorphism rs35705950, and wherein the single nucleotide polymorphism is predictive of increased survival.

21. The method of any one of embodiments 15-20, wherein the interstitial lung disease is a fibrotic interstitial lung disease.

22. The method of embodiment 21, wherein the interstitial lung disease is idiopathic pulmonary fibrosis or familial interstitial pneumonia.

23. A method of detecting the presence or absence of at least one genetic variant in a human subject, the method comprising: detecting the presence or absence of at least one genetic variant of at least one of TOLLIP, SPPL2C, and MDGA2 in a sample from the subject.

24. The method of embodiment 23, wherein the at least one genetic variant includes a genetic variant of TOLLIP.

25. The method of embodiment 23 or embodiment 24, wherein the at least one genetic variant includes a genetic variant of SPPL2C.

26. The method of any one of embodiments 23-25, wherein the at least one genetic variant includes a genetic variant of MDGA2.

27. The method of any one of embodiments 23-26, further comprising testing the sample for a genetic variant of MUC5B.

28. The method of any of embodiments 23-27, wherein the at least one genetic variant includes at least one of the genetic variants listed in Fig. 7.

29. The method of embodiment 28, wherein the at least one genetic variant includes one or more of a single nucleotide polymorphism selected from the group consisting of rs111521887, rs5743894, rs5743890, rs17690703, and rs7144383.

30. The method of embodiment 29, wherein the at least one genetic variant includes rs111521887.

31. The method of embodiment 29 or 30, wherein the wherein the at least one genetic variant includes rs5743894.

32. The method of any one of embodiments embodiment 29-31, wherein the at least one genetic variant includes rs5743890.

33. The method of any one of embodiments 29-32, wherein the at least one genetic variant includes rs17690703.

34. The method of any one of embodiments 29-33, wherein the at least one genetic variant includes rs7144383.

35. The method of any one of embodiments 29-34, further comprising testing the sample for the genetic variant rs35705950.

36. The method of any one of embodiments 22-35, wherein the subject has or is suspected of having or is at risk for developing an interstitial lung disease.

37. The method of embodiment 36, wherein the interstitial lung disease is a fibrotic interstitial lung disease or familial interstitial pneumonia.

38. The method of embodiment 37, wherein the interstitial lung disease is idiopathic pulmonary fibrosis.

39. A method of detecting the presence or absence of at least two genetic variants in a human subject having or suspected of being at risk for developing an interstitial lung disease, the method comprising: detecting the presence or absence of at least two of the genetic variants listed in Fig. 7 in a sample from the subject.

40. The method of embodiment 39, wherein the at least two genetic variants includes from two to 52 of the genetic variants listed in Fig. 7.

41. The method of embodiment 40, wherein the at least two genetic variants includes from two to 44 of the genetic variants listed in Fig. 11.

42. A method of testing for interstitial lung disease in a human subject, the method comprising: detecting a level of TOLLIP gene expression in a sample from the subject, a low level of TOLLIP gene expression relative to a control being indicative of interstitial lung disease.

43. The method of embodiment 42, wherein the level of gene expression is detected by measuring directly or indirectly TOLLIP mRNA.

44. The method of embodiment 42, wherein the level of gene expression is detected by measuring Tollip protein.

45. A method of treating a human subject having an interstitial lung disease, the method comprising: detecting a level of TOLLIP expression according to any one of embodiments 42-44; and if the subject has a low level of TOLLIP expression relative to a control, administering to the subject an amount of a Tollip agonist, Tollip

or a genetic construct expressing TOLLIP effective to treat the interstitial lung disease.

46. A kit for predicting, diagnosing, or prognosing interstitial lung disease in a human subject, the kit consisting essentially of: at least one probe or primer for detecting the presence or absence of at least one genetic variation in at least one of TOLLIP, SPPL2C, and MDGA2.

47. The kit of embodiment 46, wherein the at least one probe or primer includes probes or primers for detecting at least one genetic variation in TOLLIP.

48. The kit of embodiment 46 or 47, wherein the at least one probe or primer includes probes or primers for detecting at least one genetic variation in SPPL2C.

49. The kit of any one of embodiments 46-48, wherein the at least one probe or primer includes probes or primers for detecting at least one genetic variation in MDGA2.

50. The kit of any one of embodiments 46-49, further comprising at least one probe or primer for detecting at least one genetic variation in MUC5B.

51. The kit of any one of embodiments 46-50, wherein the genetic variation includes at least one of rs111521887, rs5743894, rs5743890, rs17690703, rs7144383, and rs35705950.

52. The kit of any one of embodiments 46-51, wherein the at least one probe or primer includes at least one probe or primer for detecting one or more of the genetic variations listed in Fig. 7.

53. A kit for predicting, diagnosing, or prognosing interstitial lung disease in a human subject, the kit comprising: at least one probe or primer for detecting the presence or absence of at least two genetic variations selected from the genetic variations listed in Fig. 7.

54. The kit of embodiment 53, wherein the kit comprises probes and/or primers for detecting the presence or absence of from two to 52 of the genetic variations listed in Fig. 7.

55. The kit of embodiment 54, wherein the kit comprises probes and/or primers for detecting the presence or absence of from two to 44 of the genetic variations listed in Fig. 11.

56. A method of determining whether a human subject has or is at risk of developing an interstitial lung disease, the method comprising detecting whether the

genome of the subject comprises at least two genetic variants selected from the group of variants listed in Fig. 7 and determining whether the subject has or is at risk of developing an interstitial lung disease, the presence of the genetic variant indicating that the subject has or is at risk of developing the interstitial lung disease.

57. The method of embodiment 56, wherein the at least two genetic variants includes from two to 52 of the genetic variants listed in Fig. 7.

58. The method of embodiment 57, wherein the at least one genetic variant includes from two to 44 of the genetic variants listed in Fig. 11.

59. The method of any one of embodiments 56-58, wherein the interstitial lung disease is a fibrotic interstitial lung disease.

60. The method of embodiment 59, wherein the interstitial lung disease is idiopathic pulmonary fibrosis or familial interstitial pneumonia.

61. A method of prognosing an interstitial lung disease in a human subject, the method comprising detecting whether the genome of the subject comprises at least two of the genetic variants listed in Fig. 7 and determining a prognosis for the subject, the presence of the genetic variant gene being prognostic of increased or decreased survival.

62. The method of embodiment 61, wherein the interstitial lung disease is a fibrotic interstitial lung disease.

63. The method of embodiment 62, wherein the interstitial lung disease is idiopathic pulmonary fibrosis or familial interstitial pneumonia.

64. A method of prognosing an interstitial lung disease in a human subject, the method comprising detecting whether the genome of the subject comprises an inversion in the 17q21.31 chromosomal region and determining a prognosis for the subject, the presence of the inversion being prognostic of increased or decreased survival.

65. A kit comprising a nucleic acid primer capable of hybridizing to a genetic variant TOLLIP nucleic acid, SPPL2C nucleic acid, or MDGA2 nucleic acid.

66. The kit of claim 65, wherein said genetic variant has been extracted from a human subject with an interstitial lung disease or is an amplification product of a nucleic acid extracted from a human subject with an interstitial lung disease.

67. The kit of claim 65 or 66, wherein said interstitial lung disease is a pulmonary fibrotic condition.

68. The kit of one of claims 65-67, further comprising a first labeled nucleic acid probe capable of hybridizing to an amplification product of said genetic variant TOLLIP nucleic acid, SPPL2C nucleic acid, or MDGA2 nucleic acid.

69. The kit of claim 68, further comprising a second labeled nucleic acid probe capable of hybridizing to an amplification product of said genetic variant TOLLIP nucleic acid, SPPL2C nucleic acid, or MDGA2 nucleic acid.

70. The kit of claim 69, wherein said first labeled nucleic acid probe comprises a first label and said additional labeled nucleic acid probe comprises a second label, wherein said first and second label are capable of fluorescence resonance energy transfer when hybridized to said genetic variant TOLLIP nucleic acid, SPPL2C nucleic acid, or MDGA2 nucleic acid.

71. An *in vitro* complex comprising a first nucleic acid probe hybridized to a genetic variant nucleic acid, said genetic variant nucleic acid comprising a genetic variant TOLLIP, SPPL2C or MDGA2 gene sequence, wherein said genetic variant nucleic acid is extracted from a human subject with an interstitial lung disease or is an amplification product of a nucleic acid extracted from a human subject with an interstitial lung disease.

72. The *in vitro* complex of claim 72, wherein said complex further comprises an second labeled nucleic acid probe hybridized to said genetic variant nucleic acid.

73. The *in vitro* complex of claim 72, wherein said first labeled nucleic acid probe comprises a first label and said second labeled nucleic acid probe comprises a second label, wherein said first and second label are capable of fluorescence resonance energy transfer.

74 An *in vitro* complex comprising a thermally stable polymerase bound to a genetic variant nucleic acid, said genetic variant nucleic acid comprising a genetic variant TOLLIP, SPPL2C or MDGA2 gene sequence, wherein said genetic variant nucleic acid is extracted from a human subject with an interstitial lung disease or is an amplification product of a nucleic acid extracted from a human subject with an interstitial lung disease.

75. The *in vitro* complex of claim 74, wherein the complex further comprises a nucleic acid primer hybridized to said genetic variant nucleic acid.

It is understood that the examples and embodiments described herein are for illustrative purposes only and that various modifications or changes in light thereof

will be suggested to persons skilled in the art and are to be included within the spirit and purview of this application and scope of the appended claims. All patents, patent applications, internet sources, and other published reference materials cited in this specification are incorporated herein by reference in their entireties. Any discrepancy between any reference material cited herein or any prior art in general and an explicit teaching of this specification is intended to be resolved in favor of the teaching in this specification. This includes any discrepancy between an art-understood definition of a word or phrase and a definition explicitly provided in this specification of the same word or phrase.

Each of the following publications is incorporated by reference in its entirety:

1. Mushiroda T, Wattanapokayakit S, Takahashi A, et al. A genome-wide association study identifies an association of a common variant in TERT with susceptibility to idiopathic pulmonary fibrosis. Journal of medical genetics 2008;45:654-6.

3. Raghu G, Brown KK, Bradford WZ, et al. A placebo-controlled trial of interferon gamma-1b in patients with idiopathic pulmonary fibrosis. The New England journal of medicine 2004;350:125-33.

4. Lederer DJ, Kawut SM, Wickersham N, et al. Obesity and primary graft dysfunction after lung transplantation: the Lung Transplant Outcomes Group Obesity Study. American journal of respiratory and critical care medicine 2011;184:1055-61.

5. Noth I, Anstrom KJ, Calvert SB, et al. A placebo-controlled randomized trial of warfarin in idiopathic pulmonary fibrosis. American journal of respiratory and critical care medicine 2012;186:88-95.

6. Raghu G, Collard HR, Egan JJ, et al. An official ATS/ERS/JRS/ALAT statement: idiopathic pulmonary fibrosis: evidence-based guidelines for diagnosis and management. American journal of respiratory and critical care medicine 2011;183:788-824.

7. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nature reviews 2010;11:499-511.

8. Gabriel SB, Schaffner SF, Nguyen H, et al. The structure of haplotype blocks in the human genome. Science (New York, NY 2002;296:2225-9.

9. R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. . 2012.

10. Han B, Eskin E. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. American journal of human genetics 2011;88:586-98.

11. Seibold MA, Wise AL, Speer MC, et al. A common MUC5B promoter polymorphism and pulmonary fibrosis. The New England journal of medicine 2011;364:1503-12.

12. American Thoracic Society. Idiopathic pulmonary fibrosis: diagnosis and treatment. International consensus statement. American Thoracic Society (ATS), and the European Respiratory Society (ERS). American journal of respiratory and critical care medicine 2000;161(2 Pt 1):646-64.

13. American Thoracic Society/European Respiratory Society International Multidisciplinary Consensus Classification of the Idiopathic Interstitial Pneumonias. This joint statement of the American Thoracic Society (ATS), and the European Respiratory Society (ERS) was adopted by the ATS board of directors, June 2001 and by the ERS Executive Committee, June 2001. American journal of respiratory and critical care medicine 2002;165(2):277-304.

14. Raghu G, Collard HR, Egan JJ, et al. An official ATS/ERS/JRS/ALAT statement: idiopathic pulmonary fibrosis: evidence-based guidelines for diagnosis and management. American journal of respiratory and critical care medicine 2011;183(6):788-824.

15. Raghu G, Brown KK, Bradford WZ, et al. A placebo-controlled trial of interferon gamma-1b in patients with idiopathic pulmonary fibrosis. The New England journal of medicine 2004;350(2):125-33.

16. Lederer DJ, Kawut SM, Wickersham N, et al. Obesity and primary graft dysfunction after lung transplantation: the Lung Transplant Outcomes Group Obesity Study. American journal of respiratory and critical care medicine 2011;184(9):1055-61.

17. Noth I, Anstrom KJ, Calvert SB, et al. A placebo-controlled randomized trial of warfarin in idiopathic pulmonary fibrosis. American journal of respiratory and critical care medicine 2012;186(1):88-95.

18. Carvalho B, Bengtsson H, Speed TP, Irizarry RA. Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data. Biostatistics (Oxford, England) 2007;8(2):485-99.

19. Carvalho BS, Louis TA, Irizarry RA. Quantifying uncertainty in genotype calls. Bioinformatics (Oxford, England) 2010;26(2):242-9.

20. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics 2007;81(3):559-75.

21. Luca D, Ringquist S, Klei L, et al. On the use of general control samples for genome-wide association studies: genetic matching highlights causal variants. American journal of human genetics 2008;82(2):453-63.

22. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS genetics 2006;2(12):e190.

23. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. Nature methods 2012;9(2):179-81.

24. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nature reviews 2010;11(7):499-511.

25. Weale ME, Depondt C, Macdonald SJ, et al. Selection and evaluation of tagging SNPs in the neuronal-sodium-channel gene SCN1A: implications for linkage-disequilibrium gene mapping. American journal of human genetics 2003;73(3):551-65.

26. Flores C, Ma SF, Maresso K, Ober C, Garcia JG. A variant of the myosin light chain kinase gene is associated with severe asthma in African Americans. Genetic epidemiology 2007;31(4):296-305.

27. A map of human genome variation from population-scale sequencing. Nature 2010;467(7319):1061-73.

28. Seibold MA, Wise AL, Speer MC, et al. A common MUC5B promoter polymorphism and pulmonary fibrosis. The New England journal of medicine 2011;364(16):1503-12.

CLAIMS

It is claimed:

1. A method of determining whether a human subject has or is at risk of developing an interstitial lung disease, the method comprising detecting whether the genome of the subject comprises a genetic variant of TOLLIP, SPPL2C or MDGA2 and determining whether the subject has or is at risk of developing an interstitial lung disease, the presence of the genetic variant indicating that the subject has or is at risk of developing the interstitial lung disease.

2. The method of claim 1, wherein the method comprises detecting whether the genome of the subject comprises a genetic variant of TOLLIP.

3. The method of claim 1 or claim 2, wherein the method comprises detecting whether the genome of the subject comprises a genetic variant of SPPL2C.

4 The method of claim 3, further comprising determining if the subject carries an H2 inversion in 17q21.31.

5. The method of claim 4, wherein the determining comprises determining if the subject comprises one or more single nucleotide polymorphisms selected from the group consisting of rs916793, rs2902662, rs17651213, and rs199448.

6. The method of any one of claims 1-5, wherein the method comprises detecting whether the genome of the subject comprises a genetic variant of MDGA2.

7. The method of any one of claims 1-6, further comprising detecting whether the genome of the subject comprises a genetic variant of MUC5B.

8. The method of any one of claims 1-7, wherein the method comprises detecting whether the genome of the subject comprises one or more genetic variants comprising a single nucleotide polymorphism selected from the group consisting of rs111521887, rs5743894, rs5743890, rs17690703, and rs7144383.

9. The method of claim 8, wherein the method comprises detecting whether the genome of the subject comprises the genetic variant comprising the single nucleotide polymorphism rs111521887.

10. The method of claim 8 or 9, wherein the method comprises detecting whether the genome of the subject comprises the genetic variant comprising the single nucleotide polymorphism rs5743894.

11. The method of any one of claims claim 8-10, wherein the method comprises detecting whether the genome of the subject comprises the genetic variant comprising the single nucleotide polymorphism rs5743890.

12. The method of any one of claims 8-11 wherein the method comprises detecting whether the genome of the subject comprises the genetic variant comprising the single nucleotide polymorphism rs17690703.

13. The method of any one of claims 8-12, wherein the method comprises detecting whether the genome of the subject comprises the genetic variant comprising the single nucleotide polymorphism rs7144383.

14. The method of any one of claims 8-13, further comprising detecting whether the genome of the subject comprises a genetic variant comprising a single polynucleotide polymorphism rs35705950.

15. The method of any one of claims 1-14, wherein the interstitial lung disease is a fibrotic interstitial lung disease.

16. The method of claim 15, wherein the interstitial lung disease is idiopathic pulmonary fibrosis or familial interstitial pneumonia.

17. A method of prognosing an interstitial lung disease in a human subject, the method comprising detecting whether the genome of the subject comprises a genetic variant of TOLLIP or SPPL2C and determining a prognosis for the subject, the presence of the genetic variant gene being prognostic of increased or decreased survival.

18. The method of claim 17, wherein the method comprises detecting whether the genome of the subject comprises a genetic variant of TOLLIP.

19. The method of claim 17 or 18, wherein the method comprises detecting whether the genome of the subject comprises a genetic variant of SPPL2C.

20. The method of claim 19, further comprising determining if the subject carries an H2 inversion in 17q21.31.

21. The method of claim 20, wherein the determining comprises determining if the subject comprises one or more single nucleotide polymorphisms selected from the group consisting of rs916793, rs2902662, rs17651213, and rs199448.

22. The method of any one of claims 17-21, further comprising detecting whether the genome of the subject comprises a genetic variant of MUC5B.

23. The method of any one of claims 17-22, wherein the genetic variant comprises a single nucleotide polymorphism selected from the group consisting of rs17690703 and rs5743890, and wherein the single nucleotide polymorphism is predictive of decreased survival.

24. The method of any one of claims 17-23, wherein the genome of the subject comprises the single nucleotide polymorphism rs35705950, and wherein the single nucleotide polymorphism is predictive of increased survival.

25. The method of any one of claims 17-24, wherein the interstitial lung disease is a fibrotic interstitial lung disease.

26. The method of claim 25, wherein the interstitial lung disease is idiopathic pulmonary fibrosis or familial interstitial pneumonia.

27. A method of detecting the presence or absence of a genetic variant in a human subject, the method comprising:

detecting the presence or absence of a genetic variant of TOLLIP, SPPL2C, or MDGA2 in a sample from the subject.

28. The method of claim 27, wherein the genetic variant is a genetic variant of TOLLIP.

29. The method of claim 27, wherein the genetic variant is a genetic variant of SPPL2C.

30. The method of claim 29, further comprising determining if the subject carries an H2 inversion in 17q21.31.

31. The method of claim 30, wherein the determining comprises determining if the subject comprises one or more single nucleotide polymorphisms selected from the group consisting of rs916793, rs2902662, rs17651213, and rs199448.

32. The method of 27, wherein the genetic variant is a genetic variant of MDGA2.

33. The method of any one of claims 27-32, further comprising detecting the presence or absence of a genetic variant of MUC5B in said sample.

34. The method of any of claims 27-33, wherein the genetic variant comprises a single nucleotide polymorphism listed in Fig. 7.

35. The method of claim 34, wherein the genetic variant comprises a single nucleotide polymorphism selected from the group consisting of rs111521887, rs5743894, rs5743890, rs17690703, and rs7144383.

36. The method of claim 35, wherein the genetic variant comprises rs111521887.

37. The method of claim 35 or 36, wherein the genetic variant comprises rs5743894.

38. The method of any one of claims 29-31, wherein the genetic variant comprises rs5743890.

33. The method of any one of claims 29-32, wherein the genetic variant comprises rs17690703.

34. The method of any one of claims 29-33, wherein the genetic variant comprises rs7144383.

35. The method of any one of claims 29-34, further comprising detecting the presence or absence of a genetic variant of MUC5B comprising rs35705950.

36. The method of any one of claims 22-35, wherein the subject has, is suspected of having, or is at risk for developing an interstitial lung disease.

37. The method of claim 36, wherein the interstitial lung disease is a fibrotic interstitial lung disease.

38. The method of claim 37, wherein the interstitial lung disease is idiopathic pulmonary fibrosis or familial interstitial pneumonia.

39. A method of detecting the presence or absence of at least two genetic variants in a human subject having, suspected of having, or at risk for developing an interstitial lung disease, the method comprising:

detecting the presence or absence of at least two of the genetic variants listed in Fig. 7 in a sample from the subject.

40. The method of claim 39, wherein the at least two genetic variants includes from two to 52 of the genetic variants listed in Fig. 7.

41. The method of claim 40, wherein the at least two genetic variants includes from two to 44 of the genetic variants listed in Fig. 11.

42. A method of testing for interstitial lung disease in a human subject, the method comprising:

detecting a level of TOLLIP gene expression in a sample from the subject, a low level of TOLLIP gene expression relative to a control being indicative of interstitial lung disease.

43. The method of claim 42, wherein the level of gene expression is detected by measuring directly or indirectly TOLLIP mRNA.

44. The method of claim 42, wherein the level of gene expression is detected by measuring Tollip protein.

45. A method of treating a human subject having an interstitial lung disease, the method comprising:

detecting a low level of TOLLIP expression relative to a control; and

administering to the subject an amount of a Tollip agonist, Tollip or a genetic construct expressing TOLLIP effective to treat the interstitial lung disease.

46. A kit for predicting, diagnosing, or prognosing interstitial lung disease in a human subject, the kit comprising:

a probe or primer capable of detecting the presence or absence of a genetic variant of TOLLIP, SPPL2C, or MDGA2.

47. The kit of claim 46, wherein the probe or primer is capable of detecting a genetic variant of TOLLIP.

48. The kit of claim 46 or 47, wherein the at least one probe or primer is capable of detecting a genetic variant of SPPL2C.

49. The kit of claim 48, further comprising at least one probe or primer that is capable of detecting an H2 inversion in 17q21.31.

50. The kit of claim 49, wherein the at least one probe or primer detects one or more single nucleotide polymorphisms selected from the group consisting of rs916793, rs2902662, rs17651213, and rs199448.

51. The kit of any one of claims 46-50, wherein the at least one probe or primer is capable of detecting a genetic variant of MDGA2.

52. The kit of any one of claims 46-51, further comprising an additional probe or primer capable of detecting at least one genetic variant of MUC5B.

53. The kit of any one of claims 46-52, wherein the genetic variant comprises rs111521887, rs5743894, rs5743890, rs17690703, rs7144383, or rs35705950.

54. The kit of any one of claims 46-53, wherein the genetic variant comprises a single nucleotide polymorphism set forth in Fig. 7.

55. A kit for predicting, diagnosing, or prognosing interstitial lung disease in a human subject, the kit comprising:

at least one probe or primer for detecting the presence or absence of at least two single nucleotide polymorphisms set forth in Fig. 7.

56. The kit of claim 55, wherein the kit comprises probes and/or primers for detecting the presence or absence of from two to 52 of the single nucleotide polymorphisms set forth in Fig. 7.

57. The kit of claim 56, wherein the kit comprises probes and/or primers for detecting the presence or absence of from two to 44 of the single nucleotide polymorphisms set forth in Fig. 11.

58. A method of determining whether a human subject has or is at risk of developing an interstitial lung disease, the method comprising detecting whether the genome of the subject comprises at least two single nucleotide polymorphisms set forth in Fig. 7 and determining whether the subject has or is at risk of developing an interstitial lung disease, the presence of the genetic variant indicating that the subject has or is at risk of developing the interstitial lung disease.

59. The method of claim 58, wherein the at least two single nucleotide polymorphisms includes from two to 52 of the single nucleotide polymorphisms set forth in Fig. 7.

60. The method of claim 59, wherein the at least two single nucleotide polymorphisms includes from two to 44 of the single nucleotide polymorphisms set forth in Fig. 11.

61. The method of any one of claims 58-60, wherein the interstitial lung disease is a fibrotic interstitial lung disease.

62. The method of claim 60, wherein the interstitial lung disease is idiopathic pulmonary fibrosis or familial interstitial pneumonia.

63. A method of prognosing an interstitial lung disease in a human subject, the method comprising detecting whether the genome of the subject comprises at least two single nucleotide polymorphisms set forth in Fig. 7 and determining a prognosis for the subject, the presence of the genetic variant gene being prognostic of increased or decreased survival.

64. The method of claim 63, wherein the interstitial lung disease is a fibrotic interstitial lung disease.

65. The method of claim 64, wherein the interstitial lung disease is idiopathic pulmonary fibrosis or familial interstitial pneumonia.

66. A method of prognosing an interstitial lung disease in a human subject, the method comprising detecting whether the genome of the subject comprises an

inversion in the 17q21.31 chromosomal region and determining a prognosis for the subject, the presence of the inversion being prognostic of increased or decreased survival.

67. A kit comprising a nucleic acid primer capable of hybridizing to a genetic variant TOLLIP nucleic acid, SPPL2C nucleic acid, or MDGA2 nucleic acid.

68. The kit of claim 67, wherein said genetic variant has been extracted from a human subject with an interstitial lung disease or is an amplification product of a nucleic acid extracted from a human subject with an interstitial lung disease.

69. The kit of claim 67 or 68, wherein said interstitial lung disease is a pulmonary fibrotic condition.

70. The kit of one of claims 67-69, further comprising a first labeled nucleic acid probe capable of hybridizing to an amplification product of said genetic variant TOLLIP nucleic acid, SPPL2C nucleic acid, or MDGA2 nucleic acid.

71. The kit of claim 70, further comprising a second labeled nucleic acid probe capable of hybridizing to an amplification product of said genetic variant TOLLIP nucleic acid, SPPL2C nucleic acid, or MDGA2 nucleic acid.

72. The kit of claim 71, wherein said first labeled nucleic acid probe comprises a first label and said additional labeled nucleic acid probe comprises a second label, wherein said first and second label are capable of fluorescence resonance energy transfer when hybridized to said genetic variant TOLLIP nucleic acid, SPPL2C nucleic acid, or MDGA2 nucleic acid.

73. An *in vitro* complex comprising a first nucleic acid probe hybridized to a genetic variant nucleic acid, said genetic variant nucleic acid comprising a genetic variant TOLLIP, SPPL2C or MDGA2 gene sequence, wherein said genetic variant nucleic acid is extracted from a human subject with an interstitial lung disease or is an amplification product of a nucleic acid extracted from a human subject with an interstitial lung disease.

74. The *in vitro* complex of claim 73, wherein said complex further comprises an second labeled nucleic acid probe hybridized to said genetic variant nucleic acid.

75. The *in vitro* complex of claim 74, wherein said first labeled nucleic acid probe comprises a first label and said second labeled nucleic acid probe comprises a second label, wherein said first and second label are capable of fluorescence resonance energy transfer.

76 An *in vitro* complex comprising a thermally stable polymerase bound to a genetic variant nucleic acid, said genetic variant nucleic acid comprising a genetic variant TOLLIP, SPPL2C or MDGA2 gene sequence, wherein said genetic variant nucleic acid is extracted from a human subject with an interstitial lung disease or is an amplification product of a nucleic acid extracted from a human subject with an interstitial lung disease.

77. The *in vitro* complex of claim 76, wherein the complex further comprises a nucleic acid primer hybridized to said genetic variant nucleic acid.

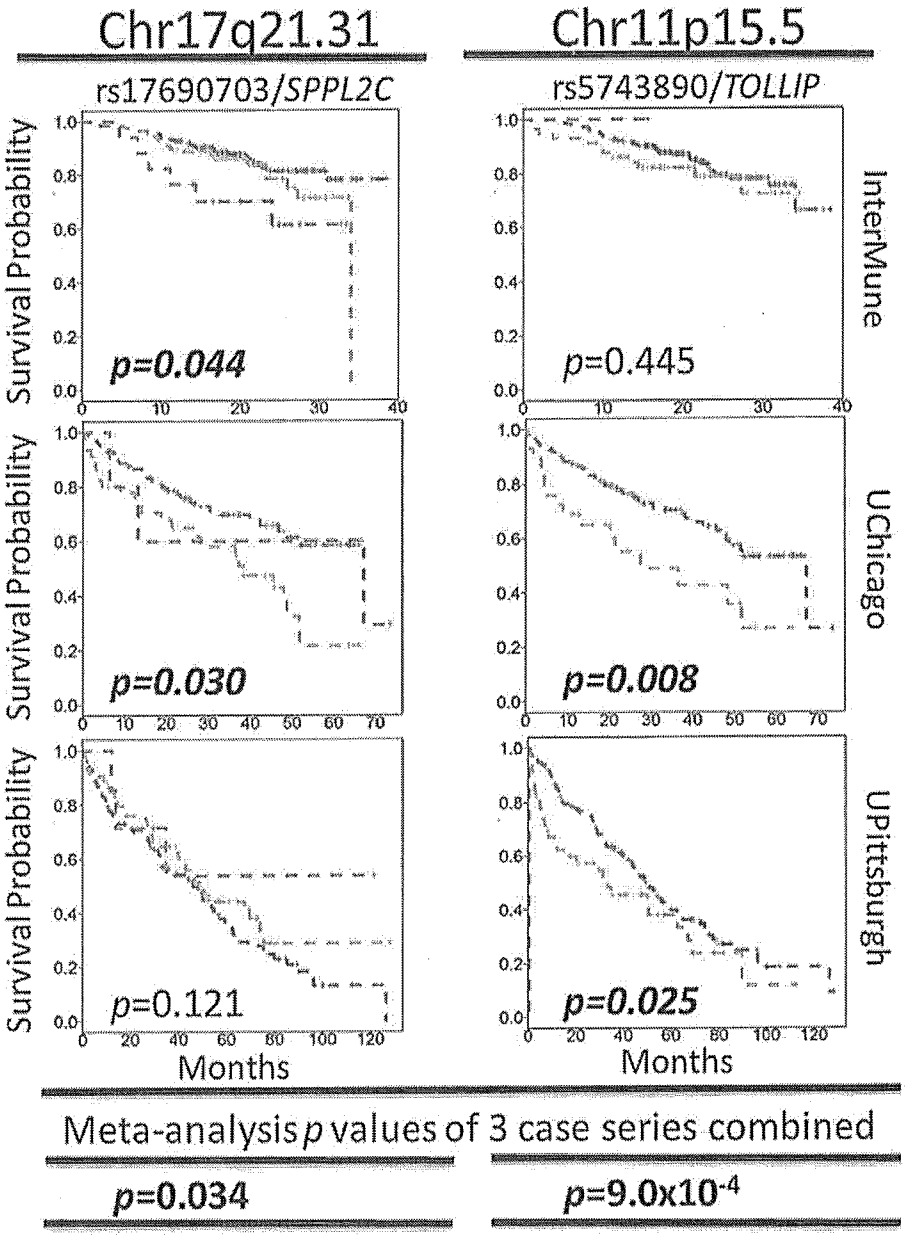
Chr17q21.31
rs17690703/SPPL2C
Chr11p15.5
rs5743890/TOLLIP
Survival Probability
p=0.044
p=0.445
InterMune
p=0.030
p=0.008
UChicago
p=0.121
p=0.025
UPittsburgh
Months
Months
Meta-analysis p values of 3 case series combined
p=0.034
p=9.0x10-4

FIG. 1

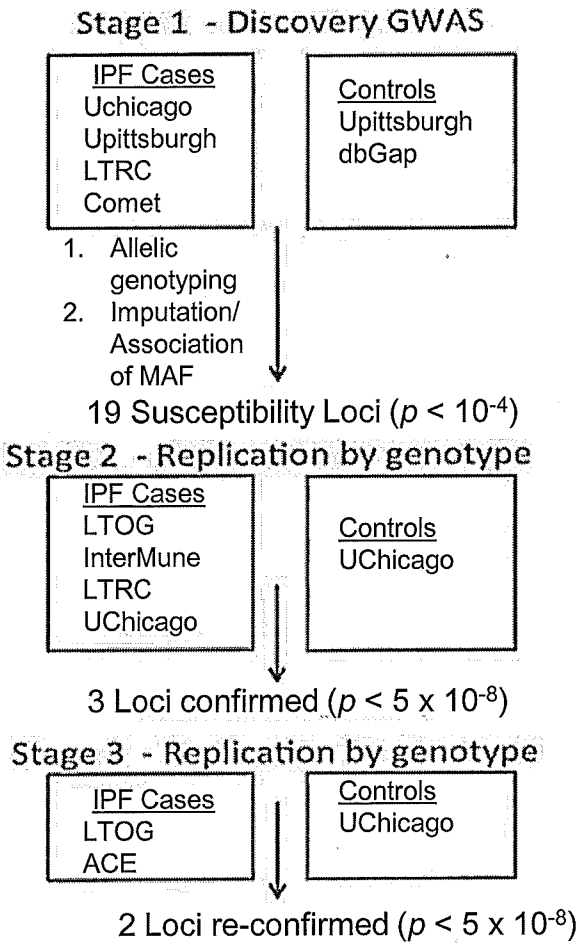
Stage 1 - Discovery GWAS
IPF Cases
Uchicago
Upittsburgh
LTRC
Comet
Controls
Upittsburgh
dbGap
1. Allelic genotyping
2. Imputation/ Association of MAF
19 Susceptibility Loci ($p < 10^{-4}$)
Stage 2 - Replication by genotype
IPF Cases
LTOG
InterMune
LTRC
UChicago
Controls
UChicago
3 Loci confirmed ($p < 5 \times 10^{-8}$)
Stage 3 - Replication by genotype
IPF Cases
LTOG
ACE
Controls
UChicago
2 Loci re-confirmed ($p < 5 \times 10^{-8}$)

FIG. 2A

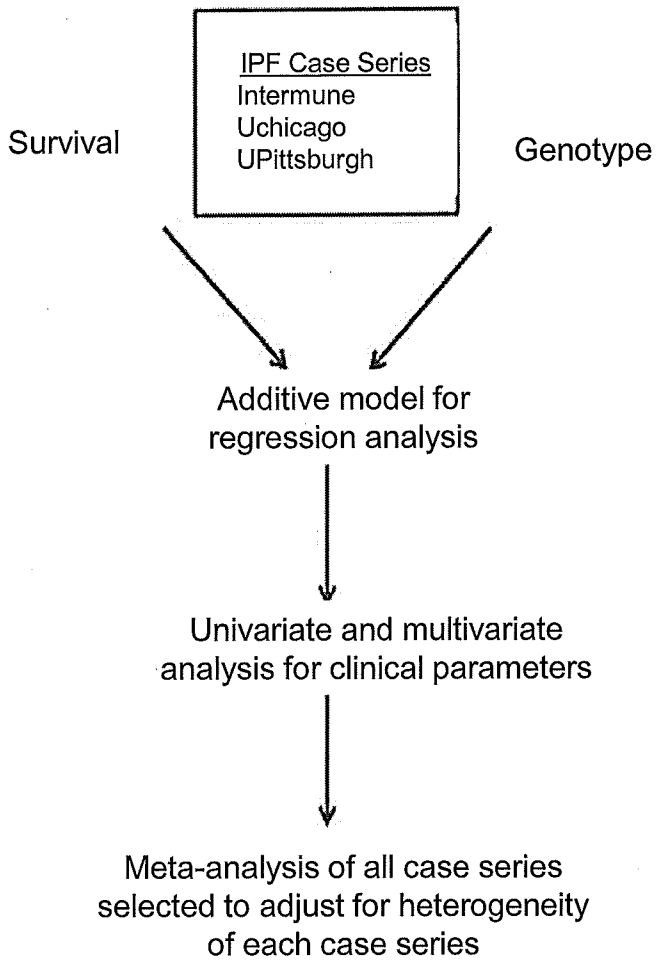
IPF Case Series
Intermune
Uchicago
UPittsburgh
Survival
Genotype
Additive model for regression analysis
Univariate and multivariate analysis for clinical parameters
Meta-analysis of all case series selected to adjust for heterogeneity of each case series

FIG. 2B

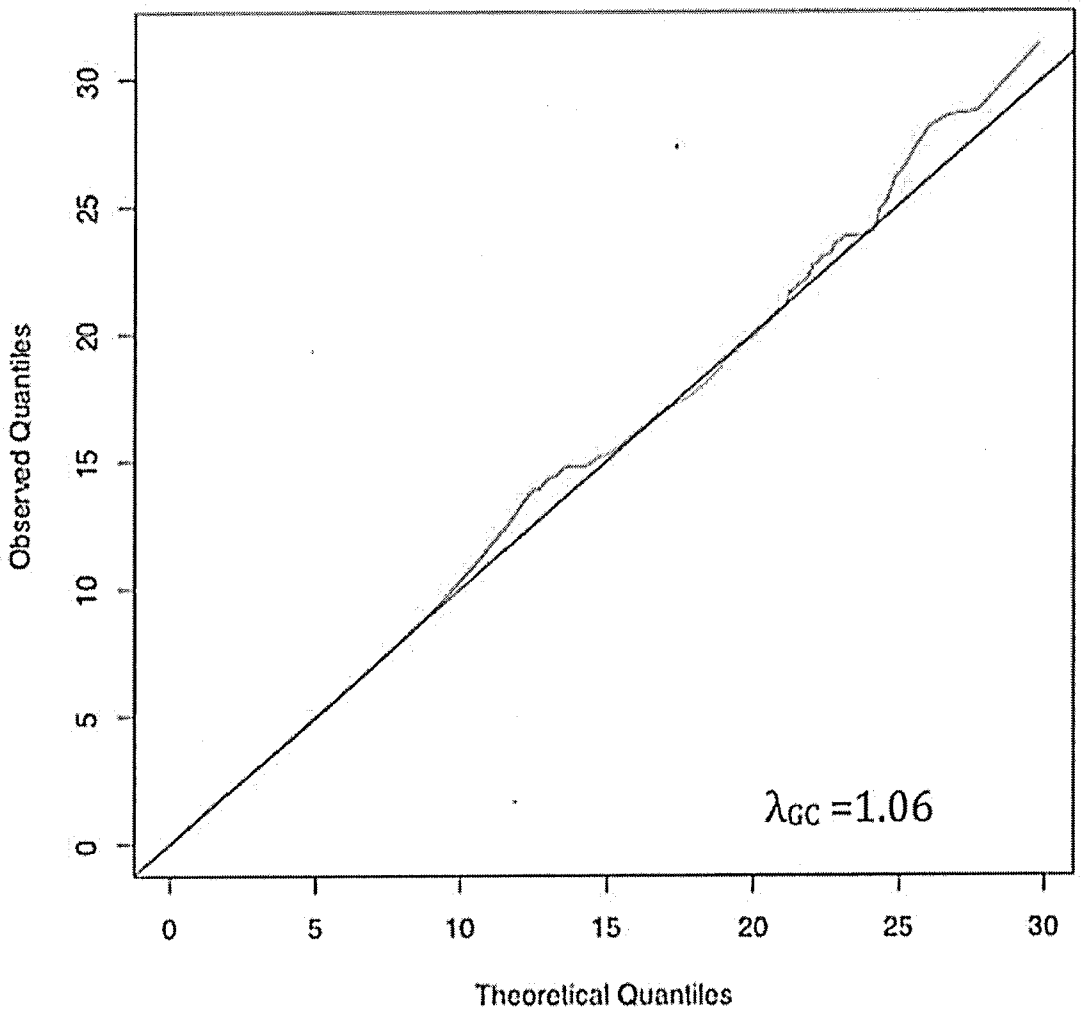
Observed Quantiles
0
5
10
15
20
25
30
Theoretical Quantiles
0
5
10
15
20
25
30
λGC =1.06

FIG. 3

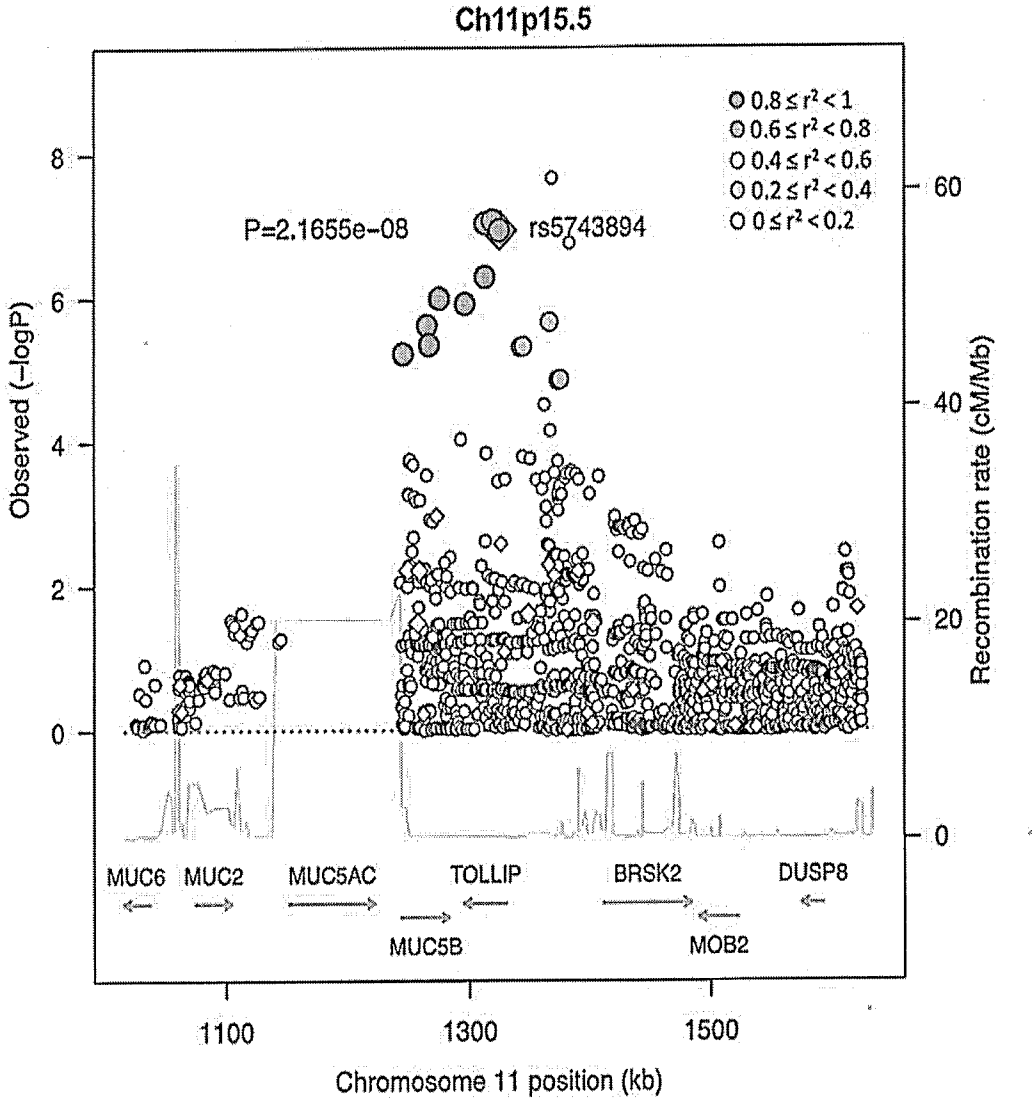
Ch11p15.5
0.8 ≤ r² < 1
0.6 ≤ r² < 0.8
0.4 ≤ r² < 0.6
0.2 ≤ r² < 0.4
0 ≤ r² < 0.2
P=2.1655e-08
rs5743894
Observed (−logP)
Recombination rate (cM/Mb)
MUC6
MUC2
MUC5AC
TOLLIP
BRSK2
DUSP8
MUC5B
MOB2
1100
1300
1500
Chromosome 11 position (kb)

FIG. 4A

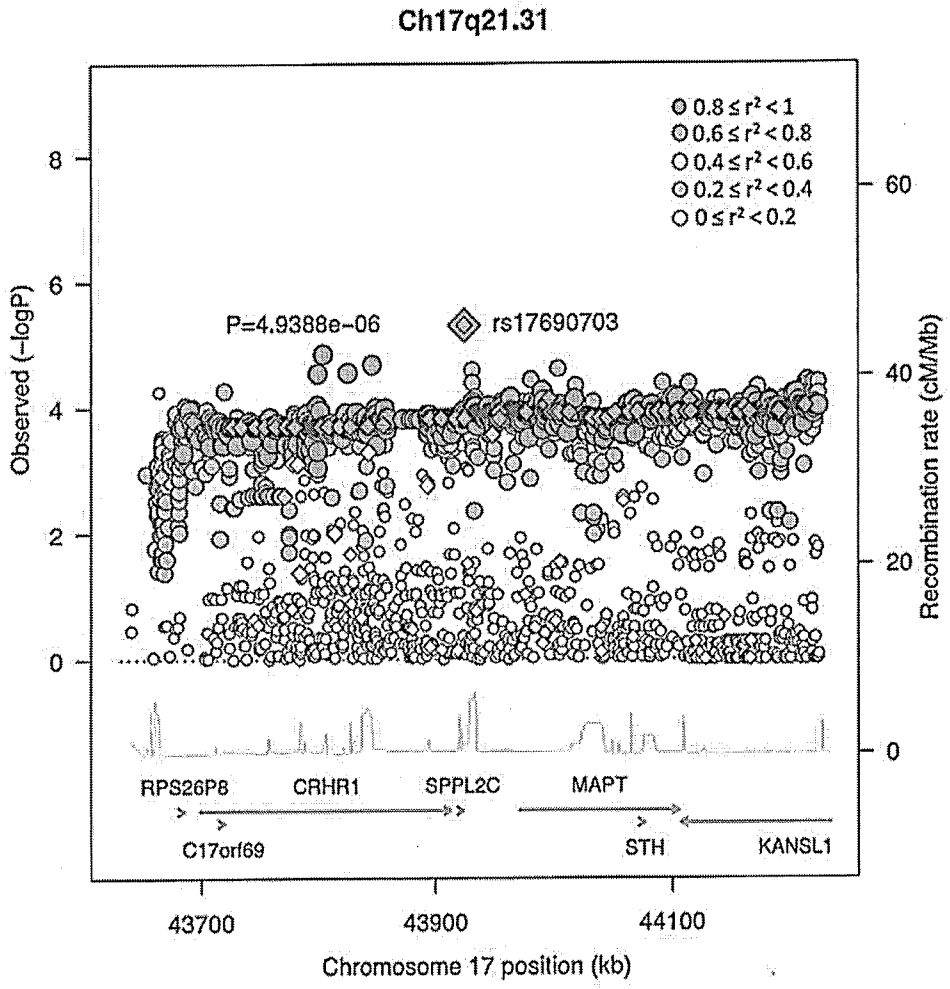
Ch17q21.31
0.8 ≤ r² < 1
0.6 ≤ r² < 0.8
0.4 ≤ r² < 0.6
0.2 ≤ r² < 0.4
0 ≤ r² < 0.2
P=4.9388e-06
rs17690703
Observed (-logP)
Recombination rate (cM/Mb)
RPS26P8
CRHR1
SPPL2C
MAPT
C17orf69
STH
KANSL1
43700
43900
44100
Chromosome 17 position (kb)

FIG. 4B

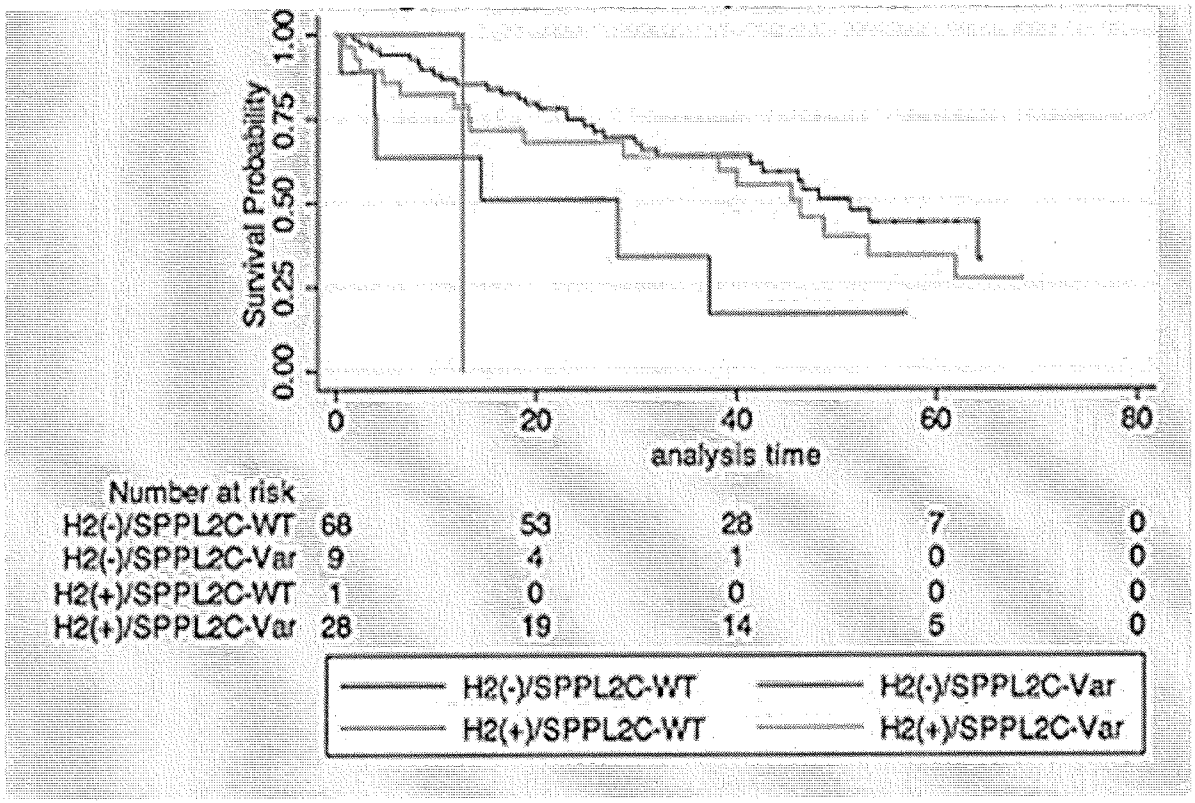
Survival Probability
0.00
0.25
0.50
0.75
1.00
0
20
40
60
80
analysis time
Number at risk
H2(-)/SPPL2C-WT 68 53 28 7 0
H2(-)/SPPL2C-Var 9 4 1 0 0
H2(+)/SPPL2C-WT 1 0 0 0 0
H2(+)/SPPL2C-Var 28 19 14 5 0
H2(-)/SPPL2C-WT
H2(-)/SPPL2C-Var
H2(+)/SPPL2C-WT
H2(+)/SPPL2C-Var

FIG 5

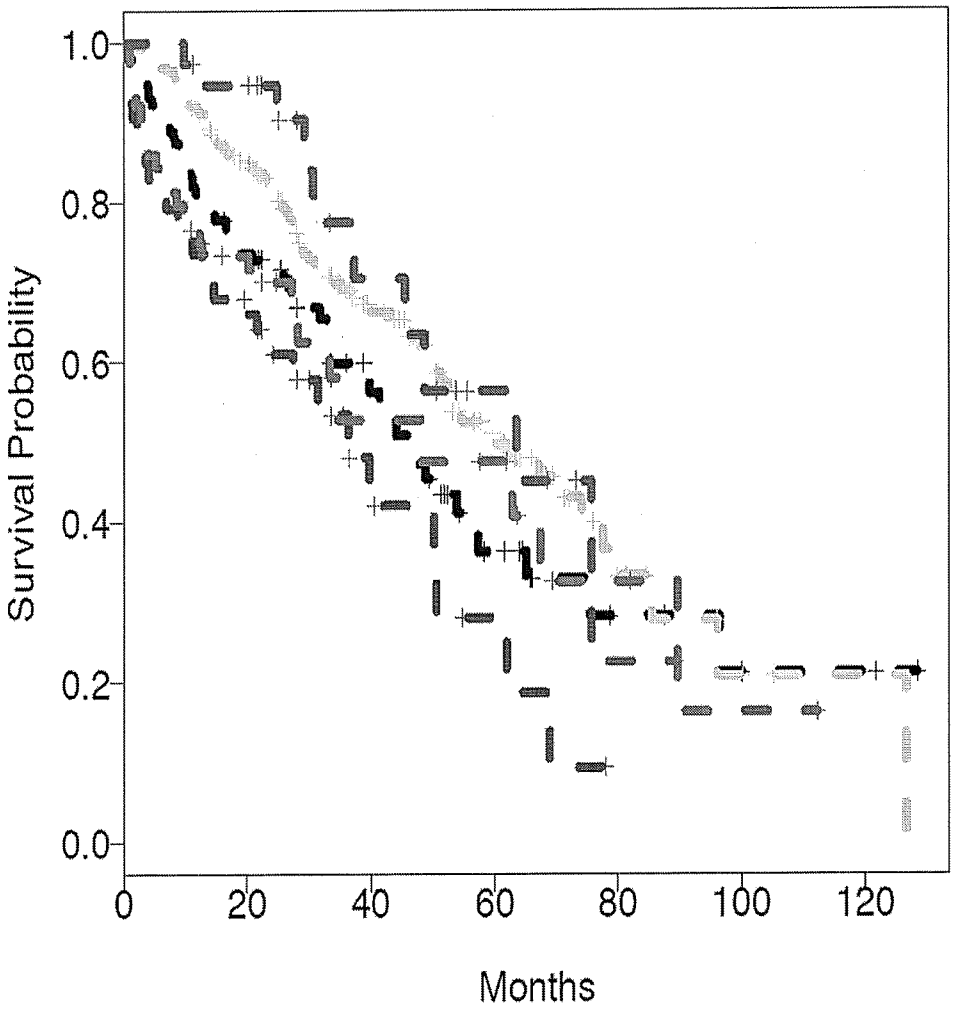
1.0
0.8
0.6
0.4
0.2
0.0
Survival Probability
0
20
40
60
80
100
120
Months

FIG. 6A

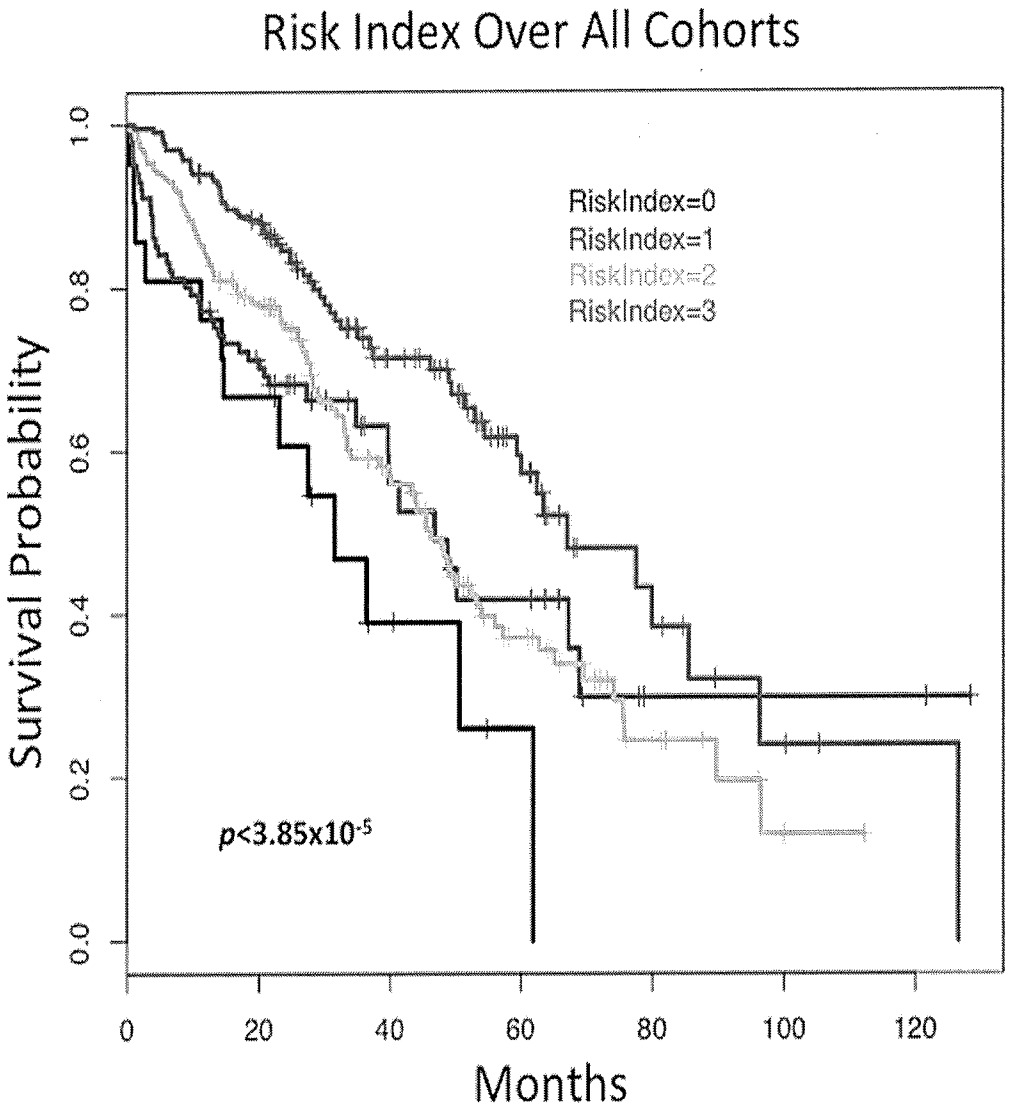
Risk Index Over All Cohorts
RiskIndex=0
RiskIndex=1
RiskIndex=2
RiskIndex=3
p<3.85x10-5
Survival Probability
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
120
Months

FIG. 6B

Table S2. List of the top 19 associated loci with susceptibility to IPF and their *p*-values from Stage 1 discovery GWAS sorted by genomic location, then by position

SNP_ID, allele	Location*	Gene/Function†	Position‡	CEU¶	EUR¶	Sources¥	IPF§	Cntrl§	HWE *p*-value	Association *p*-value
rs11810452, G	1q42.2	*SLC35F3*/intronic	234112544	0.26	0.26	1	0.44	0.53	0.794	6.31E-06
rs12750467, A	1q42.2	*SLC35F3*/intronic	234210819	0.18	0.16	0	0.22	0.16	0.316	6.31E-05
rs13034562, G	2q22	*ARHGEF4*/intronic	131713332	0.14	0.13	0	0.15	0.09	0.398	6.31E-05
rs66525751, A	2q22	*ARHGEF4*/intronic	131733483	0.09	0.10	1	0.14	0.08	1.000	7.94E-06
4_89147344, C	4q22.1	*ABCG2*/unknown	89147344	0.01	0.01	1	0.02	0.00	1.000	1.58E-06
rs10065111, T	5p15.33	--/unknown	2050328	0.22	0.21	0	0.26	0.18	1.000	1.58E-05
rs7449198, A	5p15.33	--/unknown	2062971	0.21	0.19	1	0.24	0.16	1.000	1.26E-06
rs58608739, A	5q35.2	--/unknown	175124662	0.05	0.04	1	0.01	0.04	0.631	6.31E-06
rs6862667, G	5q35.2	--/unknown	175161949	0.02	0.03	1	0.01	0.04	0.158	7.94E-05
rs55946146, G	7p22	*MAD1L1*/intronic	1866953	0.32	0.36	1	0.30	0.39	0.040	2.00E-06
rs6977733, A	7p22	*MAD1L1*/intronic	1886725	0.31	0.34	0	0.28	0.37	0.398	1.00E-05
rs3800917, A	7p22	*MAD1L1*/intronic	2167939	0.31	0.38	0	0.32	0.42	0.794	7.94E-06
rs76795398, G	7p11.2	*KC3193*/unknown	55308204	0.05	0.05	1	0.07	0.03	1.000	1.00E-05
rs79842896, C	7p11.2	--/unknown	55344174	0.04	0.04	1	0.07	0.03	1.000	7.94E-06
rs113172295, T	7q21.3	--/unknown	93979120	0.08	0.08	1	0.09	0.14	0.794	7.94E-06
rs2293739, A	7q22.1	*COL1A2*/intronic	94037778	0.09	0.12	1	0.09	0.14	0.398	3.98E-04
rs17751471, A	8p23.3	*MYOM2*/intronic	2018347	0.18	0.17	1	0.19	0.13	0.251	2.51E-05
rs6558599, G	8p23.3	*MYOM2*/intronic	2028573	0.11	0.11	0	0.14	0.08	0.794	7.94E-06
rs1425735, A	8q21.2	*PPP2R2A*/unknown	25552770	0.35	0.31	0	0.31	0.23	0.631	1.00E-04
rs4291236, C	8q22.2	*LRP12*/intronic	105550659	0.06	0.09	1	0.10	0.06	0.251	1.58E-04
rs4537272, C	8q22.2	*LRP12*/intronic	105585299	0.02	0.05	1	0.06	0.03	1.000	2.00E-06
9_25356216, T	9p21.3	--/unknown	25356216	0.00	0.01	1	0.02	0.00	1.000	7.94E-06
9_25370576, T	9p21.3	--/unknown	25370576	0.00	0.01	1	0.02	0.00	1.000	1.26E-05
rs35705950, T	11p15.5	*MUC5B*/unknown	1241221	0.10	0.10	1	0.14	0.09	1.000	6.31E-06
rs11041133, G	11p15.5	--/unknown	1291254	0.41	0.46	2	0.42	0.47	0.158	1.00E-02
rs7481967, C	11p15.5	--/unknown	1296237	0.13	0.18	2	0.14	0.14	0.100	7.94E-01

FIG. 7A

Table S2. List of the top 19 associated loci with susceptibility to IPF and their *p*-values from Stage 1 discovery GWAS sorted by genomic location, then by position

SNP_ID, allele	Location*	Gene/Function†	Position‡	CEU¶	EUR‖	Sources**	IPF††	Cntrl‡‡	HWE *p*-value	Association *p*-value
rs5744033, G	11p15.5	*TOLLIP*/UTR-3	1296263	0.13	0.25	2	0.29	0.21	1.000	1.26E-06
rs3168046, C	11p15.5	*TOLLIP*/UTR-3	1296649	0.50	0.47	2	0.52	0.47	0.025	1.00E-02
rs3829223, C	11p15.5	*TOLLIP*/intronic	1300406	0.45	0.48	2	0.44	0.49	0.200	2.51E-02
rs3750920, T	11p15.5	*TOLLIP*/cds-synon	1309956	0.45	0.46	2	0.50	0.44	0.200	5.01E-03
rs5743961, A	11p15.5	*TOLLIP*/intronic	1310714	0.06	0.06	2	0.04	0.05	0.631	2.51E-01
rs111521887, G	11p15.5	*TOLLIP*/intronic	1312706	0.22	0.20	1	0.29	0.21	0.631	5.01E-07
rs908225, T	11p15.5	*TOLLIP*/intronic	1313229	0.25	0.28	2	0.25	0.31	0.040	2.51E-03
rs5743944, A	11p15.5	*TOLLIP*/intronic	1313909	0.21	0.20	2	0.18	0.24	0.251	1.58E-04
rs5743942, C	11p15.5	*TOLLIP*/intronic	1314028	0.52	0.53	2	0.44	0.48	0.794	1.00E-01
rs5743905, G§	11p15.5	*TOLLIP*/intronic	1322713	0.06	0.06	2	0.04	0.05	0.631	2.51E-01
rs5743900, G§	11p15.5	*TOLLIP*/intronic	1323284	0.16	0.17	2	0.16	0.21	0.158	3.16E-04
rs5743899, G	11p15.5	*TOLLIP*/intronic	1323564	0.18	0.21	2	0.17	0.19	0.050	3.16E-01
rs5743894, G	11p15.5	*TOLLIP*/intronic	1324772	0.23	0.20	2	0.29	0.21	0.794	1.26E-07
rs117572864, T	11p15.5	*TOLLIP*/intronic	1324936	0.05	0.06	2	0.04	0.04	0.631	6.31E-01
rs5743890, G	11p15.5	*TOLLIP*/intronic	1325829	0.14	0.14	2	0.11	0.15	0.794	2.51E-03
rs4963062, A	11p15.5	*TOLLIP*/intronic	1326641	0.05	0.08	2	0.07	0.06	0.501	6.31E-01
rs75735056, A	11p15.5	--/unknown	1367989	0.19	0.24	1	0.34	0.24	0.126	2.00E-08
rs4898572, A	14q21.3	*MDGA2*/intronic	48005148	0.09	0.11	0	0.13	0.07	0.316	1.00E-05
rs7144383, G	14q21.3	*MDGA2*/intronic	48040375	0.09	0.11	1	0.13	0.08	0.316	2.00E-05
15_40693273, G	15q15.1	--/unknown	40693273	0.01	0.01	1	0.39	0.50	0.794	2.51E-07
rs1001528, A	15q14-q15	*IVD*/UTR-3	40713774	0.54	0.51	0	0.38	0.49	0.794	1.00E-06
rs17232873, G	15q23	*IQCH*/intronic	67769907	0.10	0.06	0	0.08	0.04	1.000	6.31E-05
rs12905544, T	15q23	*MAP2K5*/intronic	67925452	0.09	0.06	1	0.08	0.03	1.000	1.00E-06
rs17690703, T	17q21.31	*SPPL2C*/unknown, *MAPT-AS1*/unknown	43925297	0.24	0.27	0	0.18	0.26	0.398	5.01E-06
rs78795069, C	17q24.3	--/unknown	68995748	0.06	0.05	1	0.03	0.05	0.398	5.01E-04

FIG. 7B

Table S2. List of the top 19 associated loci with susceptibility to IPF and their *p*-values from Stage 1 discovery GWAS sorted by genomic location, then by position

SNP_ID, allele	Location*	Gene/Function†	Position‡	CEU§	EUR¶	Sources#	IPF§	Cntrl§	HWE *p*-value	Association *p*-value
rs721597, C	17q24.3	--/unknown	69002243	0.49	0.47	1	0.51	0.42	0.631	6.31E-06

List is compiled from the combination of genotyped or imputed top SNPs ($p<10^{-4}$) from Stage 1discovery GWAS study and tagging SNPs of *TOLLIP* selected by TagIT

*Based on Entrez Gene cytogenetic band and grouped by regions (loci)

†Abbreviations: *SLC35F3=solute carrier family 35, member; F3ARHGEF4=Rho guanine nucleotide exchange factor (GEF) 4; ABCG2=ATP-binding cassette, sub-family G (WHITE), member 2; MAD1L1=MAD1 mitotic arrest deficient-like 1 (yeast); K03193=aberrant epidermal growth factor receptor (EGFR) mRNA, complete cds; COL1A2=collagen, type I, alpha 2; MYOM2=myomesin (M-protein) 2, 165kDa; PPP2R2A=protein phosphatase 2, regulatory subunit B, alpha; LRP12=low density lipoprotein receptor-related protein 12; MUC5B=mucin 5B, oligomeric mucus/gel-forming; TOLLIP=toll interacting protein; MDGA2=MAM domain containing glycosylphosphatidylinositol anchor 2; IVD=isovaleryl-CoA dehydrogenase; IQCH=IQ motif containing H; MAP2K5=mitogen-activated protein kinase kinase 5; SPPL2C= signal peptide peptidase like 2C; MAPT-AS1=MAPT antisense RNA 1, non-coding RNA*

‡Based on GRCh37/hg19 database

§Minor allele frequency (MAF) in Utah residents with Northern and Western European ancestry from the CEPH collection (CEU) and in European ancestry individuals (CEU, FIN, GBR, IBS and TSI) in the 1000 Genomes Project Consortium. FIN=Finnish from Finland; GBR=British from England and Scotland; IBS=Iberian populations in Spain; TSI=Toscani in Italia (TSI); IPF=IPF cases; Cntrl=controls

¶0=directly genotyped on Affymetrix SNP array; 1=imputed SNP using the 1000 Genomes Project data as a reference; 2=TagIT software selected tagging SNPs

#Minor allele incorrectly listed on dbSNP

FIG. 7C

Table 1. Staged sample sizes and their corresponding studies within each stage

Stage 1: GWAS cases (n=633)		Stage 1: GWAS controls (n=1545)	
UChicago	187	UPittsburgh	103
UPittsburgh	254	dbGaP	1442
COMET	89		
LTRC	103		
Stage 2: Replication cases (n=544)		**Stage 2: Replication controls (n=687)**	
UChicago	14	UChicago	687
LTOG	174		
LTRC	42		
InterMune	314		
Stage 3: Replication cases (n=324)		**Stage 3: Replication controls (n=702)**	
LTOG	234	UChicago	702
ACE-IPF	90		

Abbreviations: UChicago=The University of Chicago; UPittsburgh=The University of Pittsburgh; COMET=Correlating Outcomes with biomedical Markers to Estimate Time-progression in IPF study; LTRC=the Lung Tissue Repository Consortium; dbGaP=the database of Genotypes and Phenotypes; LTOG=Lung Transplant Outcomes Group; InterMune=InterMune GIPF study; ACE-IPF=AntiCoagulant Effectiveness in Idiopathic Pulmonary Fibrosis

FIG. 8

Table 2. Characteristics of Idiopathic Pulmonary Fibrosis (IPF) patients in Stage 1 utilized for the discovery GWAS

Characteristics		All patients (N=633)	QC passed patients (N=542)
Age at diagnosis (yr)			
	Median	67	68
	Range	31-88	31-87
Gender			
	Male	440	385
	Female	177	157
	Unknown	16	0
Smoking status			
	Never	149	116
	Ever	338	294
	Unknown	146	132
Lung biopsies,	%	37.3	37.3
FVC % predicted,	(mean, SD)	64.9 ± 18.1	64.9 ± 17.9
D_LCO % predicted,	(mean, SD)	46.5 ± 17.8	46.5 ± 17.3

Data are presented as means ± standard deviations

FVC = forced vital capacity

D_LCO = diffusion capacity of lung for carbon monoxide

FIG. 9

Supplementary Tables and Figures

Table S1. Characteristics of Idiopathic Pulmonary Fibrosis (IPF) patients and controls by stage and by availability

Characteristics	All patients			All controls[a]		
Stage/description	1/GWAS (n=633)	2/Replication (n=544)	3/Replication (n=324)	1/GWAS (n=103)	2/Replication (n=687)	3/Replication (n=702)
No. of subjects	633[b]	370[b]/174[c]	90[b]/234[c]	103[b]	687[d]	702[d]
Age at diagnosis (yr)						
Median	67	65	68	53	56	62
Range	31-88	21-93	48-79	20-80	18-92	19-90
Gender*						
Male, n (%)	430 (71.1)	262 (70.8)	68 (75.6)	42 (41.0)	393 (57.2)	497 (70.8)
Female, n (%)	175 (28.9)	108 (29.2)	22 (24.4)	61 (59.0)	294 (42.8)	205 (29.2)
Smoking status*						
Never, n (%)	149 (30.6)	119 (32.1)	20 (22.2)	60 (58.2)	--	--
Ever, n (%)	338 (69.4)	251 (67.9)	70 (77.8)	43 (41.8)	--	--
Lung biopsies[#], %	37.3	17.6	n/a	--	--	--
FVC, % predicted (mean, SD)	64.9 ± 18.1	70.3 ± 13.6	58.4 ± 16.1	--	--	--
D_LCO, % predicted (mean, SD)	45.5 ± 17.8	47.1 ± 11.0	35.0 ± 12.7	--	--	--

[a]Individuals with genome-wide data acquired from dbGaP were not included
[b]Samples with phenotypes
[c]Samples without phenotypes
[d]Samples with partial phenotypes
*Percentage were calculated based on the number of known phenotypes
#LTOG cohort in Stage 2
Data are presented as means ± standard deviations or number (with percentage)
FVC = forced vital capacity
D_LCO = diffusion capacity of lung for carbon monoxide

FIG. 10

FIG. 11A

Table S3. List of 44 SNPs and their association *p*-values with susceptibility to IPF from Stage 1 discovery GWAS, Stage 2 replication study, and overall (Stage 1 re-genotyped and 2 combined) sorted by genomic location, then by position.

SNP[1], allele with effect	Location[2]	Gene/Function[3]	Position[4]	Types[5]	Stage 1- Discovery GWAS				Stage 2 - Replication				Overall
					IPF[6]	Cntrl[6]	OR (95%CI)[7]	*p*-value	IPF[6]	Cntrl[6]	OR (95%CI)[7]	*p*-value	*p*-value[8]
rs12750467, A	1q42.2	*SLC35F3*/intronic	234210819	0	0.22	0.16	1.55 (1.25-1.93)	6.31E-05	0.19	0.17	1.15 (0.94-1.41)	2.00E-01	2.00E-04
rs66525751, A	2q22	*ARHGEF4*/intronic	131733483	1	0.14	0.08	1.79 (1.36-2.35)	7.94E-06	0.10	0.11	0.90 (0.70-1.17)	4.00E-01	1.60E-02
4_89147344, C	4q22.1	*ABCG2*/unknown	89147344	1	0.02	0.00	7.36 (2.37-22.9)	1.58E-06	0.01	0.01	0.79 (0.31-2.00)	6.30E-01	6.90E-05
rs10065111, T	5p15.33	--/unknown	2050328	0	0.26	0.18	1.57 (1.28-1.93)	1.58E-05	0.21	0.20	1.10 (0.90-1.34)	3.20E-01	1.90E-04
rs7449198, A	5p15.33	--/unknown	2062971	1	0.24	0.16	1.65 (1.33-2.04)	1.26E-06	0.13	0.13	1.03 (0.81-1.30)	7.90E-01	4.30E-05
rs58608739, A	5q35.2	--/unknown	175124662	1	0.01	0.04	0.32 (0.18-0.56)	6.31E-06	0.04	0.03	1.11 (0.73-1.69)	6.30E-01	1.80E-02
rs6862667, G	5q35.2	--/unknown	175161949	1	0.01	0.04	0.32 (0.17-0.59)	7.94E-05	0.05	0.05	0.90 (0.63-1.29)	6.30E-01	5.60E-03
rs55948146, G	7p22	*MAD1L1*/intronic	1866953	1	0.30	0.39	0.66 (0.79-0.55)	2.00E-06	0.36	0.36	1.01 (0.86-1.20)	7.90E-01	1.50E-03
rs6977733, A	7p22	*MAD1L1*/intronic	1886725	0	0.28	0.37	0.67 (0.56-0.80)	1.00E-05	0.30	0.33	1.17 (0.99-1.40)	6.30E-02	1.30E-05
rs3800917, A	7p22	*MAD1L1*/intronic	2167939	0	0.32	0.42	0.67 (0.56-0.08)	7.94E-06	0.34	0.37	0.86 (0.73-1.02)	7.90E-02	1.40E-05
rs76795398, G	7p11.2	*K03193*/unknown	55308204	1	0.07	0.03	2.36 (1.54-3.63)	1.00E-05	0.04	0.04	0.84 (0.55-1.28)	4.00E-01	1.00E-02
rs79842896, C	7p11.2	--/unknown	55344174	1	0.07	0.03	2.49 (1.61-3.85)	7.94E-06	0.07	0.05	1.42 (1.01-1.99)	5.00E-02	2.10E-05
rs2293739, A	7q22.1	*COL1A2*/intronic	94037778	1	0.09	0.14	0.63 (0.48-0.82)	3.98E-04	0.12	0.13	1.02 (0.80-1.30)	7.90E-01	1.10E-02
rs17751471, A	8p23.3	*MYOM2*/intronic	2018347	1	0.19	0.13	1.56 (1.24-1.97)	2.51E-05	0.17	0.16	1.07 (0.86-1.32)	5.00E-01	1.70E-03
rs6558599, G	8p23.3	*MYOM2*/intronic	2028573	0	0.14	0.08	1.87 (2.48-1.41)	7.94E-06	0.12	0.11	1.12 (0.88-1.44)	4.00E-01	3.00E-04

Table S3. List of 44 SNPs and their association *p*-values with susceptibility to IPF from Stage 1 discovery GWAS, Stage 2 replication study, and overall (Stage 1 re-genotyped and 2 combined) sorted by genomic location, then by position.

SNP[1], allele with effect	Location[2]	Gene/Function[3]	Position[4]	Types[5]	Stage 1- Discovery GWAS				Stage 2 - Replication				Overall
					IPF[6]	Cntrl[6]	OR (95%CI)[7]	*p*-value	IPF[6]	Cntrl[6]	OR (95%CI)[7]	*p*-value	*p*-value[8]
rs1425735, A	8q21.2	*PPP2R2A*/unknown	25552770	0	).31	0.23	1.45 (1.76-1.20)	1.00E-04	0.29	0.28	1.06 (0.89-1.26)	5.00E-01	2.50E-03
rs4291236, C	8q22.2	*LRP12*/intronic	105550659	1	).10	0.06	1.78 (2.45-1.30)	1.58E-04	0.08	0.08	0.97 (0.73-1.31)	7.90E-01	1.50E-02
9_25356216, T	9p21.3	--/unknown	25356216	1	).02	0.00	10.27 (2.24-46.9)	7.94E-06	0.01	0.01	1.70 (0.71-4.04)	2.50E-01	2.10E-04
rs35705950, T	**11p15.5**	***MUC5B*/promoter**	**1241221**	**1**	**).14**	**0.09**	**1.57 (1.20-2.05)**	**6.31E-06**	**0.33**	**0.12**	**3.72 (3.02-4.58)**	**2.50E-40**	**2.30E-38**
rs11041133, G	11p15.5	--/unknown	1291254	2	).42	0.47	0.81 (0.96-0.68)	1.00E-02	0.41	0.45	0.82 (0.70-0.96)	1.60E-02	5.90E-04
rs7481967, C	11p15.5	--/unknown	1296237	2	).14	0.14	0.98 (1.25-0.77)	7.94E-01	0.16	0.13	1.28 (1.02-1.61)	3.20E-02	1.40E-01
rs3168046, C	11p15.5	*TOLLIP*/UTR-3	1296649	2	).52	0.47	1.24 (1.05-1.47)	1.00E-02	0.48	0.49	1.15 (0.98-1.35)	7.90E-02	2.90E-03
rs3829223, C	11p15.5	*TOLLIP*/intronic	1300406	2	).44	0.49	0.83 (0.99-0.70)	2.51E-02	0.42	0.47	0.83 (0.71-0.98)	2.50E-02	2.30E-03
rs3750920, T	11p15.5	*TOLLIP*/cds-synon	1309956	2	).50	0.44	1.25 (1.06-1.48)	5.01E-03	0.50	0.47	1.14 (0.98-1.34)	7.90E-02	1.90E-03
rs5743961, A	11p15.5	*TOLLIP*/intronic	1310714	2	).04	0.05	0.82 (0.55-1.21)	2.51E-01	0.05	0.04	1.38 (0.94-2.01)	1.00E-01	5.90E-01
rs111521887, G	**11p15.5**	***TOLLIP*/intronic**	**1312706**	**1**	**).29**	**0.21**	**1.54 (1.26-1.87)**	**5.01E-07**	**0.25**	**0.18**	**1.58 (1.29-1.94)**	**7.90E-06**	**7.10E-12**
rs908225, T	11p15.5	*TOLLIP*/intronic	1313229	2	).25	0.31	0.75 (0.91-0.62)	2.51E-03	0.26	0.30	0.79 (0.66-0.95)	1.30E-02	1.00E-04
rs5743944, A	11p15.5	*TOLLIP*/intronic	1313909	2	).18	0.24	0.69 (0.56-0.85)	1.58E-04	0.19	0.24	0.74 (0.61-0.90)	3.20E-03	2.00E-06
rs5743942, C	11p15.5	*TOLLIP*/intronic	1314028	2	).44	0.48	0.88 (0.75-1.05)	1.00E-01	0.50	0.48	0.90 (0.76-1.05)	1.60E-01	1.90E-02
rs5743905, G§	11p15.5	*TOLLIP*/intronic	1322713	2	).04	0.05	0.82 (0.55-1.21)	2.51E-01	0.05	0.04	0.75 (0.52-1.10)	1.30E-01	6.80E-01

FIG. 11B

Table S3. List of 44 SNPs and their association *p*-values with susceptibility to IPF from Stage 1 discovery GWAS, Stage 2 replication study, and overall (Stage 1 re-genotyped and 2 combined) sorted by genomic location, then by position.

SNP[1], allele with effect	Location[2]	Gene/Function[3]	Position[4]	Types[5]	Stage 1- Discovery GWAS				Stage 2 - Replication				Overall
					IPF[6]	Cntrl[6]	OR (95%CI)[7]	*p*-value	IPF[6]	Cntrl[6]	OR (95%CI)[7]	*p*-value	*p*-value[8]
rs5743900, G§	11p15.5	*TOLLIP*/intronic	1323284	2	0.16	0.21	0.69 (0.55-0.86)	3.16E-04	0.16	0.19	1.27 (1.03-1.57)	2.50E-02	5.40E-05
rs5743899, G	11p15.5	*TOLLIP*/intronic	1323564	2	0.17	0.19	0.90 (1.12-0.72)	3.16E-01	0.18	0.17	1.12 (0.90-1.37)	3.20E-01	9.10E-01
rs5743894, G	11p15.5	*TOLLIP*/intronic	1324772	2	0.29	0.21	1.58 (1.30-1.93)	1.26E-07	0.24	0.17	1.52 (1.25-1.86)	1.60E-05	2.50E-12
rs117572864, T	11p15.5	*TOLLIP*/intronic	1324936	2	0.04	0.04	1.11 (0.73-1.69)	6.31E-01	0.04	0.05	0.87 (0.59-1.29)	5.00E-01	8.30E-01
rs5743890, G	11p15.5	*TOLLIP*/intronic	1325829	2	0.11	0.15	0.68 (0.53-0.88)	2.51E-03	0.09	0.15	0.57 (0.44-0.73)	4.00E-06	7.60E-08
rs4963062, A	11p15.5	*TOLLIP*/intronic	1326641	2	0.07	0.06	1.09 (1.52-0.77)	6.31E-01	0.07	0.07	1.03 (0.76-1.42)	7.90E-01	6.30E-01
rs4898572, A	14q21.3	*MDGA2*/intronic	48005148	0	0.13	0.07	1.89 (2.52-1.42)	1.00E-05	0.10	0.09	1.09 (0.84-1.43)	5.00E-01	3.40E-04
rs7144383, G	14q21.3	*MDGA2*/intronic	48040375	1	0.13	0.08	1.82 (2.41-1.37)	2.00E-05	0.11	0.07	1.57 (1.18-2.08)	1.00E-03	3.50E-08
rs1001528, A	15q14-q15	*IVD*/UTR-3	40713774	0	0.38	0.49	0.65 (0.77-0.55)	1.00E-06	0.42	0.46	0.87 (0.74-1.03)	1.00E-01	4.80E-06
rs17232873, G	15q23	*IQCH*/intronic	67769907	0	0.08	0.04	2.16 (1.47-3.19)	6.31E-05	0.04	0.05	1.38 (0.94-2.03)	1.00E-01	7.60E-02
rs12905544, T	15q23	*MAP2K5*/intronic	67925452	1	0.08	0.03	2.53 (1.69-3.77)	1.00E-06	0.05	0.06	0.83 (0.58-1.19)	3.20E-01	9.10E-03
rs17690703, T	17q21.31	*SPPL2C*/unknown, *MAPT-AS1*/unknown	43925297	0	0.18	0.26	0.62 (0.51-0.76)	5.01E-06	0.20	0.26	0.74 (0.61-0.89)	1.60E-03	4.20E-08
rs78795069, C	17q24.3	–/unknown	68995748	1	0.03	0.06	0.52 (0.34-0.80)	5.01E-04	0.06	0.05	1.15 (0.81-1.63)	5.00E-01	1.10E-01
rs721597, C	17q24.3	–/unknown	69002243	1	0.51	0.42	1.44 (1.21-1.70)	6.31E-06	0.42	0.45	1.12 (0.95-1.31)	1.60E-01	3.50E-02

[1]SNPs with Hardy-Weinberg Equilibrium $p<10^{-3}$ in Stage 2 replication controls were excluded
[2]Based on Entrez Gene cytogenetic band and grouped by regions (loci)

FIG. 11C

Table S3. List of 44 SNPs and their association *p*-values with susceptibility to IPF from Stage 1 discovery GWAS, Stage 2 replication study, and overall (Stage 1 re-genotyped and 2 combined) sorted by genomic location, then by position.

SNP[1], allele with effect	Location[2]	Gene/Function[3]	Position[4]	Types[5]	Stage 1- Discovery GWAS				Stage 2 - Replication				Overall
					IPF[6]	Cntrl[6]	OR (95%CI)[7]	*p*-value	IPF[6]	Cntrl[6]	OR (95%CI)[7]	*p*-value	*p*-value[8]

[3]Abbreviations: *SLC35F3=solute carrier family 35, member; F3ARHGEF4=Rho guanine nucleotide exchange factor (GEF) 4; ABCG2=ATP-binding cassette, sub-family G (WHITE), member 2; MAD1L1=MAD1 mitotic arrest deficient-like 1 (yeast); K03193=aberrant epidermal growth factor receptor (EGFR) mRNA, complete cds; COL1A2=collagen, type I, alpha 2; MYOM2=myomesin (M-protein) 2, 165kDa; PPP2R2A=protein phosphatase 2, regulatory subunit B, alpha; LRP12=low density lipoprotein receptor-related protein 12; MUC5B=mucin 5B, oligomeric mucus/gel-forming; TOLLIP=toll interacting protein; MDGA2=MAM domain containing glycosylphosphatidylinositol anchor 2; IVD=isovaleryl-CoA dehydrogenase; IQCH=IQ motif containing H; MAP2K5=mitogen-activated protein kinase kinase 5; SPPL2C= signal peptide peptidase like 2; MAPT-AS1=MAPT antisense RNA 1, non-coding RNA*

[4]Based on GRCh37/hg19 database

[5]Directly genotyped on Affymetrix SNP array=0; Imputed SNP using the 1000 Genomes Project data as a reference=1; TagIT software selected tagging SNPs=2

[6]Minor allele frequency (MAF) in IPF cases and controls

[7]Odd Ratio (OR) was calculated based on minor allele frequency in cases and controls; 95% CI= 95% Confidence Interval

[§]Minor allele incorrectly listed on dbSNP

[8]SNPs with overall *p*-value < 10^{-8} are in bold

FIG. 11D

Table 3. Characteristics of Idiopathic Pulmonary Fibrosis (IPF) case series for mortality analysis

Cohort[1]		InterMune (n=314)	UChicago (n=157)	UPittsburgh (n=212)	Combined (n=683)
Average follow-up (months)		22	40	70	45
Subjects					
	Alive, n (%)	257 (81.8)	94 (61.0)	91 (42.9)	444 (65.0)
	Dead, n (%)	57 (18.2)	63 (39.0)	121 (57.1)	239 (35.0)
Age at diagnosis* (yr)					
	Median	67	69	69	69
	Range	42-79	43-93	39-84	39-93
Gender					
	Male, n (%)	228 (72.6)	118 (75.2)	143 (67.5)	449 (71.6)
	Female, n (%)	86 (27.4)	39 (24.8)	69 (32.5)	194 (28.4)
Smoking status*					
	Never, n (%)	97 (30.9)	41 (28.7)	23 (23.7)	161 (29.1)
	Ever, n (%)	217 (69.1)	102 (71.3)	74 (76.3)	393 (70.9)
#FVC % predicted,	(mean, SD)	71.56 ± 12.68	65.17 ± 18.29	65.27 ± 19.72	68.32 ± 16.39
D_LCO % predicted,	(mean, SD)	47.24 ± 8.86	48.17 ± 17.92	47.55 ± 18.77	47.61 ± 14.33

Data are presented as means ± standard deviations or number (with percentage)

[1]Lung transplant patients were censored

*Percentage were calculated based on the number of known phenotypes

#FVC = forced vital capacity, p=2.06E-05 and 4.48E-05 for InterMune vs UChicago and InterMune vs UPittsburgh

D_LCO = diffusion capacity of lung for carbon monoxide

FIG. 12

FIG. 13

Table 4. Association signals with susceptibility to IPF across stages of the six SNPs followed-up in Stage 3

			Stage 1 - discovery GWAS				Stage 2 - Replication					Stage 3 - Replication					
Gene*			MAF	MAF			MAF	MAF				MAF	MAF			Overall	
/Left	SNP_ID, allele		IPF#	Cnt#	OR		IPF#	Cnt#	OR		Combined	IPF#	Cnt#	OR		OR	Overall
/Right	with effect	Location	(542)	(542)	(95% CI)	*p*-value	(544)	(687)	(95% CI)	*p*-value	*p*-value	(324)	(702)	(95% CI)	*p*-value	(95% CI)	*p*-value
MUC5B																	
/MUC5AC					1.57				3.72					3.08		2.43	
/TOLLIP	rs35705950, T	11p15.5	0.14**	0.09	(1.20-2.05)	6.31E-06	0.33	0.12	(3.02-4.58)	2.50E-40	2.38E-38	0.31	0.14	(2.37-3.99)	5.14E-16	(2.13-2.77)	2.40E-50
TOLLIP																	
/MUC5B					1.54				1.58					1.24		1.48	
/BRSK2	rs111521887, G	11p15.5	0.29	0.21	(1.26-1.87)	5.01E-07	0.25	0.18	(1.29-1.94)	7.90E-06	7.10E-12	0.23	0.19	(0.98-1.56)	8.03E-02	(1.32-1.66)	2.20E-12
TOLLIP																	
/MUC5B					1.58				1.52					1.19		1.49	
/BRSK2	rs5743894, G	11p15.5	0.29	0.21	(1.30-1.93)	1.26E-07	0.24	0.17	(1.25-1.86)	1.60E-05	2.40E-12	0.20	0.18	(0.94-1.50)	1.57E-01	(1.33-1.68)	1.35E-12
TOLLIP																	
/MUC5B					0.68				0.57					0.60		0.61	
/BRSK2	rs5743890, G	11p15.5	0.11	0.15	(0.53-0.88)	2.51E-03	0.09	0.15	(0.44-0.73)	4.00E-06	7.68E-08	0.10	0.17	(0.45-0.80)	3.49E-04	(0.52-0.71)	3.43E-11
MDGA2																	
/NA					1.82				1.57					1.16		1.44	
/NA	rs7144383, G	14q21.3	0.13	0.08	(2.41-1.37)	2.00E-05	0.11	0.07	(1.18-2.08)	1.00E-03	3.50E-08	0.13	0.11	(0.88-1.54)	2.99E-01	(1.23-1.69)	3.71E-06
SPPL2C																	
/CRHR1					0.62				0.74					0.78		0.70	
/MAPT	rs17690703, T	17q21.31	0.18	0.26	(0.51-0.76)	5.01E-06	0.20	0.26	(0.61-0.89)	1.60E-03	4.20E-08	0.20	0.24	(0.62-0.98)	3.07E-02	(0.62-0.79)	5.70E-09

* *MUC5B*=mucin 5B, oligomeric mucus/gel-forming ; *MUC5AC*=mucin 5AC, oligomeric mucus/gel-forming; *TOLLIP*=toll interacting protein; *BRSK2*=BR serine/threonine kinase 2; *MDGA2*=MAM domain containing glycosylphosphatidylinositol anchor 2; *SPPL2C*= signal peptide peptidase like 2C; *CRHR1*=corticotropin releasing hormone receptor 1; *MAPT*=microtubule-associated protein tau; *NA*=not available

#MAF=minor allele frequency; Cnt=control; *p*-value in a joint analysis of 2 and 3 Stages were in bold.

**Shown here is the MAF from imputation; physical genotyped MAF=0.34. The MAF discrepancy may due to limited available information in reference 1000 Genomes Project

Table 54. SNP effects on mortality in each case series and overall case samples.

SNP_ID, allele with effect	InterMune (n=314)[2]	UChicago (n=157)[1]	UPittsburgh (n=212)[1]	Combined (n=683)[1]
rs35705950, T				
Alive, *Genotype count (MAF)	73/165/17 (0.390)	17/63/8 (0.449)	16/25/5 (0.380)	106/253/30 (0.400)
Dead, Genotype count (MAF)	19/36/1 (0.339)	23/33/1 (0.307)	51/61/8 (0.321)	93/130/10 (0.322)
Odd Ratio (95% CI)	0.79 (0.46-1.35)	0.39 (0.20-0.75)	0.79 (0.50-1.25)	0.65 (0.49-0.87)
p-value	0.387	**0.005**	0.317	**0.003**
rs111521887, G				
Alive, Genotype count (MAF)	126/112/19 (0.292)	36/39/9 (0.339)	22/23/1 (0.272)	184/174/29 (0.300)
Dead, Genotype count (MAF)	29/26/2 (0.263)	28/22/1 (0.235)	70/43/4 (0.218)	127/91/7 (0.233)
Odd Ratio (95% CI)	0.91 (0.57-1.47)	0.94 (0.54-1.61)	0.79 (0.48-1.31)	0.82 (0.62-1.07)
p-value	0.712	0.808	0.364	0.141
rs5743894, G				
Alive, Genotype count (MAF)	132/111/14 (0.270)	52/24/7 (0.229)	22/22/2 (0.283)	206/157/23 (0.263)
Dead, Genotype count (MAF)	29/28/0 (0.246)	35/16/1 (0.173)	73/43/4 (0.213)	137/87/5 (0.212)
Odd Ratio (95% CI)	0.92 (0.56-1.52)	0.99 (0.55-1.79)	0.74 (0.45-1.22)	0.82 (0.61-1.08)
p-value	0.749	0.976	0.235	0.159
rs5743890, G				
Alive, Genotype count (MAF)	212/44/1 (0.089)	79/13/0 (0.071)	39/7/0 (0.076)	330/64/1 (0.084)
Dead, Genotype count (MAF)	44/13/0 (0.114)	41/15/0 (0.134)	91/29/1 (0.128)	176/57/1 (0.126)
Odd Ratio (95% CI)	1.25 (0.62-2.50)	2.26 (1.00-5.14)	1.38 (0.72-2.62)	1.54 (1.06-2.24)
p-value	0.536	0.050	0.332	**0.025**
rs7144383, G				
Alive, Genotype count (MAF)	202/55/0 (0.107)	66/24/0 (0.133)	40/5/1 (0.076)	308/84/1 (0.109)
Dead, Genotype count (MAF)	48/9/0 (0.079)	47/11/1 (0.110)	89/28/2 (0.134)	184/48/3 (0.115)
Odd Ratio (95% CI)	0.73 (0.34-1.57)	0.99 (0.48-2.05)	1.34 (0.73-2.45)	1.11 (0.77-1.59)
p-value	0.420	0.973	0.351	0.583
rs17690703, T				
Alive, Genotype count (MAF)	176/70/11 (0.179)	67/20/3 (0.144)	28/16/2 (0.217)	271/106/16 (0.176)
Dead, Genotype count (MAF)	32/19/6 (0.272)	30/24/2 (0.250)	84/33/3 (0.163)	146/76/11 (0.210)
Odd Ratio (95% CI)	1.63 (1.03-2.57)	1.88 (1.03-3.43)	0.72 (0.44-1.19)	1.20 (0.91-1.58)
p-value	**0.035**	**0.040**	0.204	0.200

*Genotype count is presented as homozygous common/heterozygous/homozygous rare; Minor allele frequency (MAF) is derived from chromosome number in each group

[1]Lung transplant individuals were censored; Time at risk was defined as the interval between data of enrollment in each cohort and date of the last follow-up, lung transplant, or death; HR=Hazard Ratio; 95% CI= 95% Confidence Intervals; Mortality association *p*-values ≤0.05 are in bold

FIG. 14

Table 5. Summaries of univariate Cox regression analysis for mortality in InterMune, UChicago, UPittsburgh, and Meta-analysis

	InterMune (n=314)[1]		UChicago (n=157)[1]		UPittsburgh (n=212)[1]		Meta-analysis (n=683)[1]	
SNP_ID, allele with effect	HR (95% CI)	*p*-value	HR (95% CI)	*p*-value	HR (95% CI)	*p*-value	HR (95% CI)	*p*-value
rs111521887, G	0.94 (0.61-1.46)	0.795	0.91 (0.59-1.39)	0.650	0.88 (0.63-1.23)	0.458	0.90 (0.72-1.13)	0.382
rs5743894, G	0.96 (0.61-1.51)	0.844	0.95 (0.59-1.50)	0.810	0.85 (0.61-1.18)	0.319	0.90 (0.71-1.13)	0.358
rs5743890, G	1.28 (0.68-2.39)	0.445	2.18 (1.22-3.88)	**0.008**	1.60 (1.06-2.40)	**0.025**	1.65 (1.23-2.21)	**0.0009**
rs7144383, G	0.78 (0.38-1.60)	0.500	1.06 (0.59-1.89)	0.844	1.10 (0.75-1.63)	0.626	1.03 (0.77-1.38)	0.849
rs17690703, T	1.50 (1.01-2.24)	**0.044**	1.57 (1.05-2.34)	**0.030**	0.76 (0.54-1.07)	0.121	1.20 (0.75-1.94)[2]	**0.034**[2]

[1]Lung transplant individuals were censored. *p*-values were derived from unadjusted univariate analyses

[2]Results for a random model are reported based on the evidence suggesting study heterogeneity. $p<0.05$ is in bold

Time at risk was defined as the interval between date of enrollment in each cohort and date of the last follow-up, lung transplant, or death; HR=Hazard Ratio; 95% CI= 95% Confidence Intervals

Dotted lines delimit the loci

FIG. 15

Table S5. Summaries of univariate and multivariate Cox regression analysis for mortality in InterMune, UChicago, UPittsburgh, and Meta-analysis.

Case series (sample size)	Analysis[2]	Values	rs11521887, G	*rs5743890, G	rs5743894, G	rs7144383, G	*rs17690703, T
InterMune (n=314)[1]	unadjusted	HR (95% CI)	0.94 (0.61-1.46)	1.28 (0.68-2.39)	0.96 (0.60-1.51)	0.78 (0.38-1.60)	1.50 (1.01-2.24)
		p-value	0.795	0.445	0.844	0.500	**0.044**
	adjusted	HR (95% CI)	0.92 (0.59-1.43)	1.60 (0.85-3.00)	0.98 (0.61-1.56)	0.77 (0.37-1.57)	1.50 (0.97-2.16)
		p-value	0.716	0.145	0.927	0.464	0.069
UChicago (n=157)[1]	unadjusted	HR (95% CI)	0.91 (0.59-1.39)	2.18 (1.22-3.88)	0.95 (0.59-1.50)	1.06 (0.59-1.89)	1.57 (1.05-2.34)
		p-value	0.650	**0.008**	0.810	0.844	**0.030**
	adjusted	HR (95% CI)	0.96 (0.59-1.57)	2.36 (1.22-4.55)	1.11 (0.67-1.84)	1.83 (0.98-3.43)	1.67 (0.98-2.83)
		p-value	0.873	**0.010**	0.693	0.057	0.057
UPittsburgh (n=212)[1]	unadjusted	HR (95% CI)	0.88 (0.63-1.23)	1.60 (1.06-2.40)	0.85 (0.61-1.18)	1.10 (0.75-1.63)	0.76 (0.54-1.07)
		p-value	0.458	**0.025**	0.319	0.626	0.121
	adjusted	HR (95% CI)	1.06 (0.76-1.49)	1.55 (0.98-2.44)	1.03 (0.74-1.45)	1.00 (0.66-1.54)	0.83 (0.57-1.21)
		p-value	0.734	0.059	0.853	0.989	0.334
Meta-analysis (n=683)[1]	unadjusted	HR (95% CI)	0.90 (0.72-1.13)	1.65 (1.23-2.21)	0.90 (0.71-1.13)	1.03 (0.77-1.38)	1.20 (0.75-1.94)[3]
		p-value	0.382	**0.0009**	0.358	0.849	**0.034**[3]
	adjusted	HR (95% CI)	1.00 (0.79-1.26)	1.73 (1.25-2.38)	1.03 (0.81-1.32)	1.11 (0.81-1.52)	1.23 (0.80-1.92)[3]
		p-value	0.972	**0.0009**	0.785	0.517	0.112[3]

*SNP_ID and alleles associated with mortality

[1]Lung transplant individuals were censored; Time at risk was defined as the interval between data of enrollment in each cohort and date of the last follow-up, lung transplant, or death; HR=Hazard Ratio; 95% CI= 95% Confidence Intervals; Mortality association *p*-values <0.05 are in bold

[2]Univariate analyses as "unadjusted"; Multivariate analyses adjusting for all recorded covariates (i.e. age, gender, tobacco history, forced vital capacity (FVC) percent predicted, diffusing capacity of carbon monoxide (D_LCO) percent predicted, and recruitment center) that maintained *p*-value<0.1 in regression models as "adjusted"

[3]Results for a random model are reported based on the evidence suggesting study heterogeneity.

FIG. 16

Table S6. Summaries of Kaplan-Meier survival analysis in InterMune, UChicago, and UPittsburgh cohorts

	InterMune (n=314)			UChicago (n=154)[1]			UPittsburgh (n=167)[1]		
SNP_ID/allele	coe	HR (95% CI)	*p*-value	coe	HR (95% CI)	*p*-value	coe	HR (95% CI)	*p*-value
rs111521887/ C>G	-0.12	0.89 (0.58-1.37)	5.92E-01	-0.41	0.66 (0.42-1.06)	8.28E-02	-0.15	0.86 (0.62-1.20)	3.66E-01
rs17690703/ C>T	0.42	1.53 (1.03-2.27)	**3.26E-02**	0.45	1.57 (1.04-2.36)	**2.98E-02**	-0.19	0.83 (0.59-1.17)	2.79E-01
rs35705950/ G>T	-0.37	0.69 (0.43-1.12)	1.33E-01	-0.72	0.49 (0.30-0.79)	**3.35E-03**	-0.26	0.77 (0.57-1.04)	9.17E-02
rs5743890/ A>G	0.30	1.34 (0.73-2.46)	3.36E-01	0.77	2.17 (1.19-3.93)	**9.11E-03**	0.56	1.76 (1.17-2.64)	**6.33E-03**
rs5743894/ A>G	-0.12	0.89 (0.56-1.40)	6.12E-01	-0.26	0.77 (0.47-1.25)	2.83E-01	-0.19	0.83 (0.60-1.15)	2.54E-01
rs7144383/ A>G	-0.27	0.76 (0.37-1.55)	4.53E-01	-0.15	0.86 (0.47-1.60)	6.44E-01	0.15	1.17 (0.79-1.72)	4.33E-01

[1]Lung transplant individuals are excluded

Survival time at risk was defined as the interval between data of enrollment in each cohort and date of the last follow-up or death. The heterogeneity of the Kaplan-Meier survival curves as a function of genotypes for each SNP was assessed by the log-rank test. Pooled hazard ratio (HR) estimates were obtained by including study as a stratification variable. $p<0.05$ is in bold

FIG. 17

Table S7. Predictors of survival in IPF patients identified using a univariant Cox model

	InterMune (n=314)				UChicago (n=154)				UPittsburgh (n=167)			
	coe	HR (95% CI)	p_SNP	p_Covar	pLogtest	coe	HR (95% CI)	p_SNP	p_Covar	pLogtest	coe	HR (95% CI)	p_SNP	p_Covar	pLogtest
Age	0.03	1.03 (0.97-NA)	6.76E-02	NA	6.67E-02	0.03	1.03 (0.97-NA)	2.24E-02	NA	2.21E-02	0.03	1.03 (0.97-NA)	2.62E-02	NA	2.58E-02
rs111521887	-0.12	0.89 (1.13-0.97)	5.90E-01	6.75E-02	1.59E-01	-0.43	0.65 (1.54-0.96)	6.23E-02	1.61E-02	1.02E-02	-0.17	0.84 (1.18-0.97)	3.17E-01	1.85E-02	3.97E-02
rs17690703	0.42	1.52 (0.66-0.97)	3.70E-02	7.22E-02	1.83E-02	0.52	1.68 (0.6-0.97)	1.87E-02	1.93E-02	6.17E-03	-0.12	0.89 (1.13-0.97)	4.94E-01	3.53E-02	5.93E-02
rs35705950	-0.42	0.66 (1.53-0.97)	9.02E-02	5.94E-02	5.24E-02	-0.73	0.48 (2.07-0.97)	3.31E-03	2.53E-02	8.36E-04	-0.24	0.79 (1.27-0.97)	1.23E-01	1.52E-02	1.25E-02
rs5743890	0.40	1.49 (0.67-0.96)	1.99E-01	4.70E-02	8.26E-02	0.84	2.31 (0.43-0.97)	6.26E-03	4.11E-02	4.16E-03	0.47	1.61 (0.62-0.98)	2.66E-02	8.09E-02	5.53E-03
rs5743894	-0.11	0.90 (1.12-0.97)	6.39E-01	6.93E-02	1.64E-01	-0.32	0.73 (1.38-0.97)	1.91E-01	2.14E-02	3.72E-02	-0.20	0.82 (1.22-0.97)	2.25E-01	2.12E-02	3.53E-02
rs7144383	-0.22	0.80 (1.25-0.97)	5.38E-01	7.54E-02	1.55E-01	-0.12	0.89 (1.13-0.97)	7.01E-01	3.89E-02	1.04E-01	0.12	1.13 (0.88-0.97)	5.28E-01	2.76E-02	6.42E-02
Gender	0.11	1.11 (0.90-NA)	7.23E-01	NA	7.23E-01	0.37	1.45 (0.69-NA)	2.55E-01	NA	2.52E-01	0.43	1.54 (0.65-NA)	3.08E-02	NA	2.95E-02
rs111521887	-0.12	0.89 (1.12-0.90)	5.98E-01	7.33E-01	8.18E-01	-0.44	0.64 (1.55-0.58)	6.48E-02	1.56E-01	7.57E-02	-0.19	0.83 (1.2-0.61)	2.75E-01	1.62E-02	3.46E-02
rs17690703	0.43	1.54 (0.65-0.87)	3.21E-02	6.35E-01	9.12E-02	0.44	1.55 (0.65-0.67)	3.65E-02	2.28E-01	4.43E-02	-0.19	0.83 (1.21-0.65)	2.84E-01	3.49E-02	5.83E-02
rs35705950	-0.37	0.69 (1.44-0.93)	1.32E-01	8.11E-01	3.14E-01	-0.73	0.48 (2.07-0.69)	3.37E-03	2.72E-01	7.23E-03	-0.26	0.77 (1.3-0.65)	8.83E-02	3.34E-02	2.34E-03
rs5743890	0.30	1.35 (0.74-0.89)	3.30E-01	6.94E-01	5.84E-01	0.73	2.08 (0.48-0.72)	1.65E-02	3.37E-01	2.15E-02	0.47	1.60 (0.62-0.71)	2.76E-02	9.72E-02	6.41E-03
rs5743894	-0.12	0.89 (1.13-0.90)	6.07E-01	7.16E-01	8.23E-01	-0.29	0.75 (1.33-0.52)	2.45E-01	9.21E-02	1.27E-01	-0.22	0.80 (1.25-0.62)	1.91E-01	2.08E-02	3.45E-02
rs7144383	-0.27	0.76 (1.31-0.90)	4.57E-01	7.31E-01	7.11E-01	-0.14	0.87 (1.15-0.68)	6.62E-01	2.45E-01	4.54E-01	0.11	1.11 (0.90-0.67)	5.84E-01	4.51E-02	9.71E-02
Tobacco	-0.12	0.88 (1.13-NA)	3.82E-01	NA	3.81E-01	0.28	1.32 (0.76-NA)	3.72E-01	NA	3.70E-01					
rs111521887	-0.11	0.89 (1.12-1.13)	6.17E-01	3.94E-01	6.01E-01	-0.45	0.64 (1.57-0.60)	6.92E-02	1.34E-01	7.63E-02					
rs17690703	0.43	1.54 (0.65-1.14)	3.15E-02	3.42E-01	6.58E-02	0.36	1.44 (0.7-0.69)	1.23E-01	2.65E-01	1.43E-01					
rs35705950	-0.37	0.69 (1.45-1.15)	1.28E-01	3.11E-01	1.98E-01	-0.80	0.45 (2.23-0.66)	2.06E-03	2.01E-01	6.33E-03					
rs5743890	0.31	1.36 (0.73-1.14)	3.17E-01	3.58E-01	4.14E-01	0.75	2.11 (0.47-0.82)	1.86E-02	5.48E-01	3.66E-02					
rs5743894	-0.11	0.90 (1.12-1.13)	6.37E-01	3.93E-01	6.09E-01	-0.32	0.72 (1.38-0.67)	2.02E-01	2.37E-01	2.28E-01					
rs7144383	-0.30	0.74 (1.35-1.14)	4.16E-01	3.48E-01	4.86E-01	-0.02	0.98 (1.02-0.71)	9.60E-01	2.89E-01	5.53E-01					
FVC % predicted	-0.03	0.97 (1.03-NA)	8.56E-03	NA	7.82E-03	-0.06	0.94 (1.06-NA)	2.30E-10	NA	6.90E-10	-0.02	0.98 (1.02-NA)	1.35E-05	NA	1.25E-05
rs111521887	-0.14	0.87 (1.15-1.03)	5.37E-01	7.96E-03	2.35E-02	-0.49	0.61 (1.64-1.05)	5.99E-02	5.94E-08	2.34E-07	-0.03	0.97 (1.03-1.02)	8.50E-01	1.95E-05	8.97E-05
rs17690703	0.40	1.49 (0.67-1.03)	4.94E-02	1.07E-02	4.14E-03	0.44	1.55 (0.65-1.05)	5.41E-02	2.06E-09	3.40E-10	-0.40	0.67 (1.49-1.03)	2.97E-02	2.12E-06	4.51E-06
rs35705950	-0.40	0.67 (1.49-1.03)	1.07E-01	7.08E-03	8.79E-03	-1.03	0.36 (2.80-1.06)	1.56E-04	9.89E-11	2.74E-11	-0.13	0.88 (1.14-1.02)	4.06E-01	4.66E-05	9.86E-05
rs5743890	0.36	1.43 (0.70-1.03)	2.45E-01	7.18E-03	1.62E-02	0.54	1.71 (0.58-1.06)	7.99E-02	6.87E-09	4.09E-09	0.86	2.35 (0.42-1.03)	1.13E-04	8.19E-07	4.80E-08
rs5743894	-0.12	0.89 (1.12-1.03)	6.13E-01	8.36E-03	2.51E-02	-0.19	0.83 (1.21-1.06)	4.64E-01	1.48E-08	7.25E-08	-0.07	0.93 (1.07-1.02)	6.79E-01	1.98E-05	7.79E-05
rs7144383	-0.31	0.73 (1.36-1.03)	3.94E-01	7.84E-03	1.94E-02	0.31	1.37 (0.73-1.06)	3.36E-01	2.19E-10	2.81E-09	0.18	1.20 (0.83-1.02)	3.88E-01	1.48E-05	7.76E-05
DLCO % predicted	-0.04	0.96 (1.04-NA)	3.16E-02	NA	3.09E-02	-0.05	0.95 (1.05-NA)	1.00E-08	NA	8.70E-09	-0.03	0.97 (1.03-NA)	3.05E-07	NA	2.89E-07
rs111521887	-0.09	0.91 (1.10-1.04)	6.79E-01	3.27E-02	8.94E-02	-0.46	0.63 (1.58-1.05)	8.48E-02	6.11E-07	2.66E-06	0.00	1.00 (1.00-1.03)	9.78E-01	9.67E-07	5.38E-06
rs17690703	0.39	1.47 (0.68-1.04)	5.37E-02	4.35E-02	1.39E-02	0.52	1.69 (0.59-1.06)	5.14E-02	4.25E-08	7.94E-08	-0.11	0.90 (1.12-1.03)	5.54E-01	1.80E-07	5.00E-07
rs35705950	-0.30	0.74 (1.35-1.03)	2.18E-01	5.42E-02	5.48E-02	-1.14	0.32 (3.14-1.06)	2.74E-04	3.86E-09	2.93E-09	-0.13	0.87 (1.14-1.03)	3.96E-01	3.31E-07	1.62E-06
rs5743890	0.43	1.54 (0.65-1.04)	1.67E-01	1.95E-02	3.95E-02	0.83	2.29 (0.44-1.05)	1.61E-02	3.22E-07	1.57E-07	0.56	1.75 (0.57-1.03)	1.11E-02	2.63E-07	1.03E-07
rs5743894	-0.08	0.92 (1.08-1.04)	7.32E-01	3.31E-02	9.20E-02	-0.24	0.79 (1.27-1.06)	3.55E-01	5.11E-07	1.77E-06	-0.13	0.97 (1.03-1.03)	8.54E-01	5.00E-07	2.55E-06
rs7144383	-0.19	0.82 (1.22-1.04)	5.94E-01	3.70E-02	8.57E-02	0.43	1.53 (0.65-1.06)	2.14E-01	2.91E-09	1.29E-08	0.09	1.09 (0.92-1.03)	6.82E-01	6.77E-07	3.72E-06

FIG. 18

Table S8. Predictors of survival in IPF patients identified using a multivariate analysis of covariance

InterMune (n=314)

	coe	HR (95% CI)	*p*-SNP	*p*-Gender	*p*-FVC	*p*-DLCO	*p*-Logtest
Gender	0.03	1.03 (0.97-0.97)	9.29E-01	NA	2.63E-02	1.21E-01	2.58E-02
FVC % predicted	-0.03	0.97 (NA-0.97)	2.63E-02	9.29E-01	NA	1.21E-01	2.58E-02
DLCO % predicted	-0.03	0.97 (0.97-NA)	1.21E-01	9.29E-01	2.63E-02	NA	2.58E-02
rs111521887	-0.11	0.90 (0.97-0.97)	6.26E-01	9.42E-01	2.54E-02	1.32E-01	4.77E-02
rs17690703	0.37	1.45 (0.97-0.98)	6.82E-02	8.48E-01	3.21E-02	1.68E-01	1.18E-02
rs35705950	-0.35	0.70 (0.97-0.98)	1.58E-01	9.12E-01	2.01E-02	2.06E-01	2.78E-02
rs5743890	0.45	1.57 (0.97-0.97)	1.45E-01	9.02E-01	2.42E-02	8.34E-02	2.55E-02
rs5743894	-0.18	0.92 (0.97-0.97)	7.21E-01	9.20E-01	2.62E-02	1.31E-01	5.03E-02
rs7144383	-0.25	0.78 (0.97-0.97)	4.93E-01	9.75E-01	2.35E-02	1.42E-01	4.46E-02

UChicago (n=154)

	coe	HR (95% CI)	*p*-SNP	*p*-Gender	*p*-FVC	*p*-DLCO	*p*-Logtest
Gender	0.29	1.33 (0.97-0.96)	4.82E-01	NA	1.63E-02	2.01E-04	6.23E-09
FVC % predicted	-0.03	0.97 (NA-0.96)	1.63E-02	4.82E-01	NA	2.01E-04	6.23E-09
DLCO % predicted	-0.04	0.96 (0.97-NA)	2.01E-04	4.82E-01	1.63E-02	NA	6.23E-09
rs111521887	-0.50	0.60 (0.98-0.96)	6.75E-02	2.28E-01	6.25E-02	1.33E-03	2.11E-06
rs17690703	0.49	1.63 (0.98-0.96)	7.46E-02	5.21E-01	4.52E-02	3.89E-04	3.51E-08
rs35705950	-1.21	0.30 (0.97-0.96)	1.11E-04	5.13E-01	1.07E-02	1.52E-04	8.54E-10
rs5743890	0.74	2.10 (0.98-0.96)	3.32E-02	6.35E-01	3.61E-02	3.37E-04	1.53E-07
rs5743894	-0.23	0.80 (0.97-0.96)	4.01E-01	3.05E-01	1.57E-02	2.63E-04	8.39E-07
rs7144383	0.43	1.54 (0.97-0.96)	2.10E-01	5.95E-01	1.80E-02	7.71E-05	4.59E-09

FIG. 19A

Table S8. Predictors of survival in IPF patients identified using a multivariate analysis of covariance

	UPittsburgh (n=167)						
	coe	HR (95% CI)	*p*-SNP	*p*-Gender	*p*-FVC	*p*-DLCO	*p*-Logtest
Gender	0.94	2.57 (0.99-0.96)	2.68E-05	NA	3.48E-02	2.51E-07	2.23E-10
FVC % predicted	-0.01	0.99 (NA-0.96)	3.48E-02	2.68E-05	NA	2.51E-07	2.23E-10
DLCO % predicted	-0.04	0.96 (0.99-NA)	2.51E-07	2.68E-05	3.48E-02	NA	2.23E-10
rs111521887	0.10	1.10 (0.99-0.96)	5.73E-01	2.20E-05	3.60E-02	5.09E-07	1.79E-09
rs17690703	-0.23	0.80 (0.99-0.96)	2.30E-01	2.28E-05	2.07E-02	4.36E-07	1.41E-10
rs35705950	-0.05	0.95 (0.99-0.96)	7.66E-01	2.86E-05	3.94E-02	2.62E-07	9.55E-10
rs5743890	0.63	1.87 (0.99-0.97)	7.54E-03	3.20E-04	8.17E-03	1.58E-06	3.08E-11
rs5743894	0.07	1.07 (0.99-0.96)	6.71E-01	2.94E-05	3.40E-02	3.16E-07	1.27E-09
rs7144383	-0.02	0.98 (0.99-0.97)	9.27E-01	5.50E-05	4.66E-02	6.21E-07	3.28E-09

FIG. 19B

Cluster	p val	Gene	biotype	Description
4.6	**5.2**	SNORA24	snoRNA	small nucleolar RNA, H/ACA box 24
4.6	4.9	SNHG8	lincRNA	small nucleolar RNA host gene 8 (non-protein coding)
4.6	3.8	PRSS12	prot_coding	protease, serine, 12 (neurotrypsin, motopsin)
4.6	3.2	NDST3	prot_coding	N-deacetylase/N-sulfotransferase 3
4.6	2.3	SEC24D	prot_coding	SEC24 family, member D (S. cerevisiae)
4.6	2.3		lincRNA	
11.1	**5.1**	TOLLIP	prot_coding	toll interacting protein
11.1	5.1		antisense	
11.1	4.4	BRSK2	prot_coding	BR serine/threonine kinase 2
11.1	4.1		antisense	
11.1	4	MUC5B	prot_coding	mucin 5B, oligomeric mucus/gel-forming
11.1	2.5	SCT	prot_coding	secretin
11.1	2.3	CDHR5	prot_coding	cadherin-related family member 5
9.2	**4.9**		snoRNA	Small nucleolar RNA U13
18.4	**4.6**		lincRNA	
22.3	4.3	EIF4NIF1	prot_coding	euk translation initiation factor 4E nuclear import factor1
22.3	**3.7**	DRG1	prot_coding	developmentally regulated GTP binding protein 1
22.3	3.4		snRNA	U6 spliceosomal RNA
22.3	3.2		snRNA	U6 spliceosomal RNA
22.3	2.7	SFI1	prot_coding	Sfi1 homolog, spindle assembly associated (yeast)
22.3	2.4	PISD	prot_coding	phosphatidylserine decarboxylase
6.1	**4**	PEX3	prot_coding	peroxisomal biogenesis factor 3

FIG. 20

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2014/014395

A. CLASSIFICATION OF SUBJECT MATTER

C12Q 1/68(2006.01)i, C12N 15/11(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
C12Q 1/68; A61K 39/395; C12N 15/63; C40B 30/04; C12N 15/11

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
Korean utility models and applications for utility models
Japanese utility models and applications for utility models

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
eKOMPASS(KIPO internal) & keywords: genetic variant, TOLLIP, SPPL2C, MDGA2, H2 inversion, 17q21.31, SNP, MUC5B, interstitial lung disease

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 2011-0217315 A1 (SCHWARTZ et al.) 8 September 2011 See abstract; paragraphs [0028], [0093].	27-29,32,33(1),67
A		30,31,38(1),33(2),46-50,68,69,73-77
Y	ZHU et al., `Tollip, an intracellular trafficking protein, is a novel modulator of the transforming growth factor-β signaling pathway` The Journal of Biological Chemistry, Vol.287, No.47, pp.39653-39663 (16 November 2012) See abstract.	27,28,33(1),67
Y	MARTIN et al., `Regulated intramembrane proteolysis of Bri2 (Itm2b) by ADAM10 and SPPL2a/SPPL2b` The Journal of Biological Chemistry, Vol.283, No.3, pp.1644-1652 (18 January 2008) See abstract.	27,29,33(1),67

☒ Further documents are listed in the continuation of Box C. ☒ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents,such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
29 April 2014 (29.04.2014)	**01 May 2014 (01.05.2014)**

Name and mailing address of the ISA/KR
International Application Division
Korean Intellectual Property Office
189 Cheongsa-ro, Seo-gu, Daejeon Metropolitan City, 302-701,
Republic of Korea
Facsimile No. +82-42-472-7140

Authorized officer
KIM, Seung Beom
Telephone No. +82-42-481-3371

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2014/014395

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☒ Claims Nos.: 1-26,36(2),37(2),38(2),42-45,58-66
because they relate to subject matter not required to be searched by this Authority, namely:
Claims 1-26, 36(2), 37(2), 38(2), 42-45, and 58-66 directed to a treatment method of the human body by therapy, as well as diagnostic methods, and thus relate to a subject matter which this International Searching Authority is not required, under Article 17(2)(a)(i) of the PCT and Rule 39.1(iv) of the Regulations under the PCT, to search.

2. ☒ Claims Nos.: 9,16,26,34(1),35(1),36(1),37(1),35(2),36(2),37(2),38(2),39-41,54-65,71,72
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:
Claims 9, 16, 26, 35(1), 36(1), 37(2), 38(2), 71, and 72 are unclear since they refer to claims which are not searchable due to not being drafted in accordance with Rule 6.4(a); Claims 34(1), 36(2), 37(2), 38(2), 39-41, and 54-65 contain a reference to the description and/or drawings. According to PCT Rule 6.2(a), claims should not contain such references except where absolutely necessary, which is not the case.

3. ☒ Claims Nos.: 6-8,10-15,22-25,34(1),37(1),34(2),35(2),36(2),51-54,70
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

1. ☐ As all required addtional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying an additional fees, this Authority did not invite payment of any additional fees.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest

☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.
☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.
☐ No protest accompanied the payment of additional search fees.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2014/014395

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 2011-0311512 A1 (HAKONARSON et al.) 22 December 2011 See abstract; claims 1-2. *Note: For claims 33(1)-38(1) and 33(2)-38(2), said claims were renumbered by this authority because claims 33-38 were found twice.	27,32,33(1),67

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.
PCT/US2014/014395

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2011-0217315 A1	08/09/2011	CA 2787994 A1	04/08/2011
		EP 2529033 A1	05/12/2012
		EP 2529033 A4	17/07/2013
		MX 2012008730 A	17/12/2012
		WO 2011-094345 A1	04/08/2011
US 2011-0311512 A1	22/12/2011	AU 2010-313759 A1	20/05/2010
		CA 2766246 A1	20/05/2010
		EP 2376655 A2	19/10/2011
		JP 2012-511895 A	31/05/2012
		WO 2010-057112 A2	20/05/2010