Gene expression profiling: can we identify the right target genes?

J.E. Loyd

ABSTRACT: **Gene expression profiling allows the simultaneous monitoring of the transcriptional behaviour of thousands of genes, which may potentially be involved in disease development. Several studies have been performed in idiopathic pulmonary fibrosis (IPF), which aim to define genetic links to the disease in an attempt to improve the current understanding of the underlying pathogenesis of the disease and target pathways for intervention.**

Expression profiling has shown a clear difference in gene expression between IPF and normal lung tissue, and has identified a wide range of candidate genes, including those known to encode for proteins involved in extracellular matrix formation and degradation, growth factors and chemokines.

Recently, familial pulmonary fibrosis cohorts have been examined in an attempt to detect specific genetic mutations associated with IPF. To date, these studies have identified families in which IPF is associated with mutations in the gene encoding surfactant protein C, or with mutations in genes encoding components of telomerase.

Although rare and clearly not responsible for the disease in all individuals, the nature of these mutations highlight the importance of the alveolar epithelium in disease pathogenesis and demonstrate the potential for gene expression profiling in helping to advance the current understanding of idiopathic pulmonary fibrosis.

KEYWORDS: Familial, idiopathic pulmonary fibrosis, microarray assays

CORRESPONDENCE
J.E. Loyd
Division of Allergy
Pulmonary and Critical Care
Dept of Medicine
Room T1218
Vanderbilt Medical Center North
Vanderbilt University School of Medicine
1161 21st Avenue S
Nashville
TN 37232-2650
USA
Fax: 1 6153437448
E-mail: jim.loyd@vanderbilt.edu

Idiopathic pulmonary fibrosis (IPF) is a progressive and mostly lethal disease with a largely unknown underlying pathogenesis [1]. Although published studies are currently limited, gene expression profiling is now being applied to IPF in an attempt to identify genetic links that might improve the current understanding of its underlying molecular mechanisms [2, 3]. High-throughput gene expression profiling technologies, such as microarray assays, have been widely used in biomedical research in recent years and allow the simultaneous monitoring of the transcriptional behaviour of thousands of genes [2–6]. Historically, microarray assays were considered to be limited by issues arising from data management and analysis, with the identification of data that were both biologically meaningful and statistically significant being considered a particular challenge [2, 4]. However, the application of more rigorous statistical benchmarking to microarray data has enabled advances to be made in the current understanding of the responses of a variety of cellular systems to stimuli, the classification and prognosis of certain cancers, and the underlying pathogenesis of a variety of disease states including IPF [6, 7].

MICROARRAY ASSAYS IN IPF

Using oligonucleotide microarrays, Zuo et al. [3] demonstrated that gene expression patterns clearly distinguish normal lung tissue from samples of patients with histologically proven pulmonary fibrosis (usual interstitial pneumonia; UIP). Over 7,000 genes were identified and major differences in expression levels were seen; with 164 genes being classified as highly informative, considerably more than would be expected (fig. 1). This abundance of informative genes beyond expected levels suggests that the differences in gene expression observed between normal and fibrotic lungs were biologically meaningful [3]. The study presented a large number of candidate genes, and many genes that were significantly increased in fibrotic lungs were found to encode proteins associated with extracellular matrix (ECM) formation and degradation, as well as proteins expressed in smooth muscle [3]. Of these, matrilysin (matrix metalloproteinase (MMP)-7), a metalloprotease not previously associated with pulmonary fibrosis, was found to be the most distinctive between fibrotic and normal lungs, along with immunohistochemistry demonstrating increased expression

of matrilysin protein in fibrotic lungs [3]. A later study by PARDO *et al.* [8] also implicated MMP-7 in IPF.

Oligonucleotide arrays identified the gene for the protein osteopontin, a cell adhesion and migration molecule that has been implicated in a number of pathological conditions, as another gene that significantly distinguishes IPF from normal lungs [8]. Microarray data suggested a significant interaction between osteopontin and MMP-7 and both proteins co-localised to alveolar epithelial cells in IPF lungs. More recently, SELMAN *et al.* [9] compared the gene expression patterns from patients with IPF, hypersensitivity pneumonitis (HP) and nonspecific interstitial pneumonia (NSIP) using custom oligonucleotide microarrays and demonstrated statistically significant gene expression signatures, which characterised HP and IPF (fig. 2). SELMAN *et al.* [9] also demonstrated that the profile of gene expression in HP was characterised by the expression of genes that are functionally associated with inflammation, T-cell activation and immune responses. However, the profile of gene expression in IPF was characterised by the expression of genes encoding tissue remodelling, and those considered to be epithelial and myofibroblast specific [9]. Using these gene expression signatures it was possible to classify NSIP, a histological pattern that is often

difficult to differentiate consistently from HP and IPF, into IPF-like or HP-like gene expression patterns, or into a pattern which resembled neither signature, probably representing idiopathic NSIP [9].

Gene expression profiling can also distinguish between familial and idiopathic forms of idiopathic interstitial pneumonia (IIP). YANG *et al.* [10] profiled RNA from the lungs of patients with sporadic IIP, familial IIP and normal control subjects on a whole human genome oligonucleotide microarray. Significant transcriptional differences were found between familial and sporadic IIP within the same functional categories as transcripts that distinguish IIP from normal samples [10]. Relevant categories included chemokines and growth factors and their receptors, complement components, and genes associated with cell proliferation and death. Given that the study by YANG *et al.* [10] used a completely different platform and completely different samples, yet showed that the same functional categories of genes were affected as demonstrated in earlier studies, these data support the reproducibility of gene expression array [11].
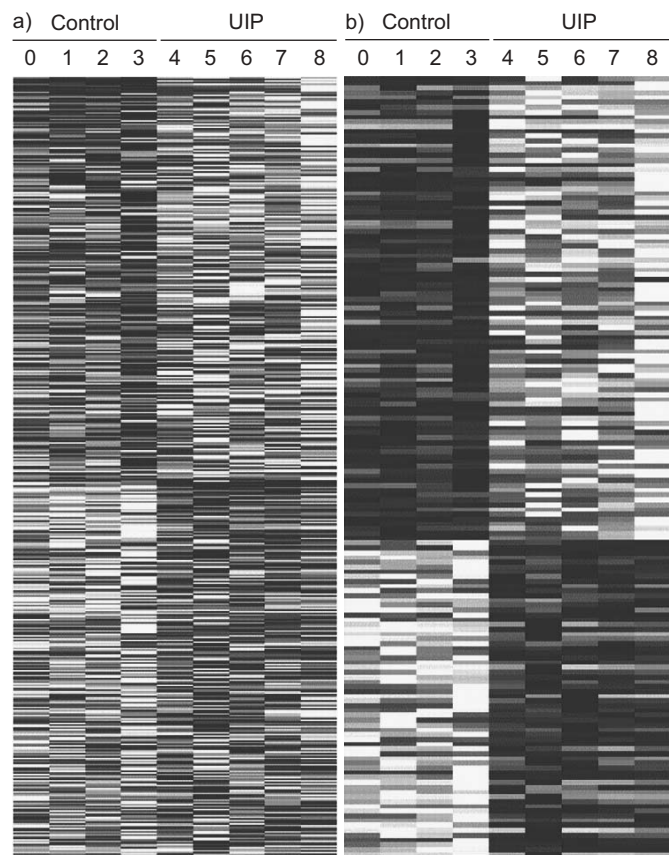

FIGURE 1. Gene expression infogram for a) all 7,129 genes and b) the 164 most informative genes showing differences between fibrotic and normal lung. To eliminate outlier effect, genes were normalised to a range of 0–1, *i.e.* the maximum value for every gene was set to 1, the minimum value was set to 0 and the rest of the values were linearly fitted to this range. A paler colour indicates maximal expression and a darker colour is minimal expression. UIP: usual interstitial pneumonia. Reproduced from [3] with permission from the publisher.
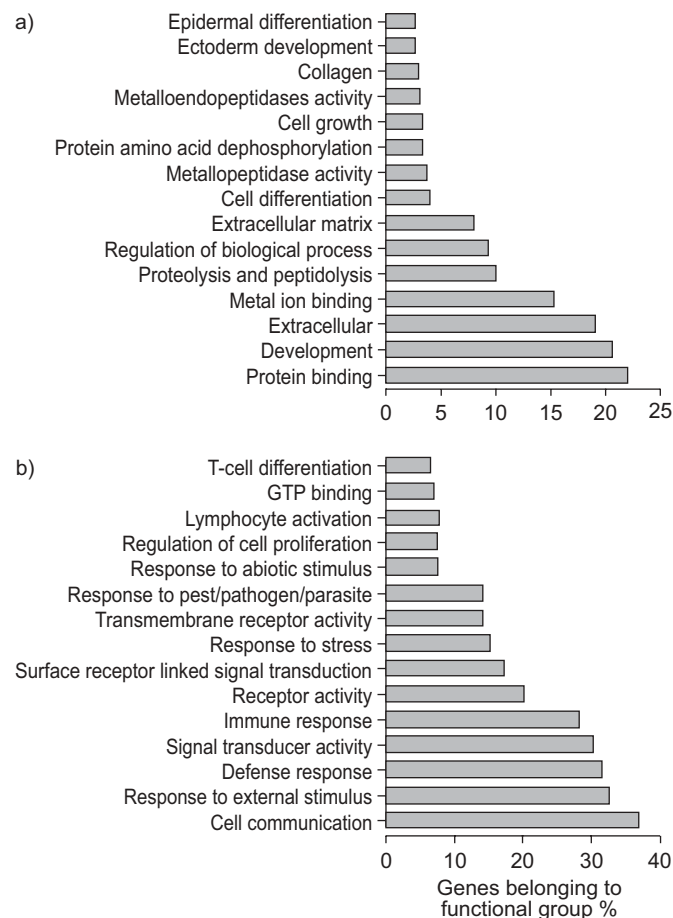

FIGURE 2. The distribution of differentially expressed genes among functional categories. a) Genes that had locus-link information and were significantly increased in idiopathic pulmonary fibrosis (IPF) compared with hypersensitivity pneumonitis (HP). b) Genes that were increased in HP compared with IPF. Note the large percentage of genes belonging to the distinct functional annotations. All enrichments are statistically significant (Fisher's exact score p<0.05). Reproduced from [9] with permission from the publisher.

Taken together, the results from these studies demonstrate the value of gene expression signatures in the classification of interstitial lung diseases, and in understanding their aetiology and underlying pathogenetic mechanisms. However, the large number of potential candidate genes identified by gene microarrays makes it very difficult to elucidate which are the most important. In addition, a number of questions arise concerning the relevance of these data. Which cell type or types contribute to the signal? Are the pivotal pathogenetic cell types dying or dead? Is the signal from a primary pathogenic process or is it part of the host response? Is it part of the initiation or propagation of the pathological response, and does gene expression correlate with protein levels? However, despite theoretical concerns the field is clearly moving forward and providing both broad and specific candidates in interstitial lung disease.

FAMILIAL IPF

Familial IPF occurs as an autosomal dominant disorder with variable penetrance. The clinical characteristics of familial IPF are virtually indistinguishable from sporadic IPF [12], although there is evidence to suggest that familial cases may be diagnosed at a younger age [13]. The incidence of familial IPF has been stated as 0.5–2% of cases of IPF [14]; however, experience in the present author's centre suggests that the actual figure may be considerably higher, ~20% [15]. The present authors and co-workers have collected data from >200 hundred families at Vanderbilt University School of Medicine (Nashville, TN, USA) alone, and ⩾500 hundred families have been identified across the USA between three collaborating centres (including Duke University, National Jewish Health/University of Colorado (both Denver, CO, USA), and Vanderbilt University). Most of these families exhibited vertical transmission from parent to child, indicative of a single dominant gene. Individuals from 111 families, containing members with familial IPF, showed a similar age at diagnosis, a significant risk among siblings and evidence consistent with an autosomal dominant pattern of inheritance [16]. A substantial proportion (45%) of families had phenotypically heterogeneous IIP (*i.e.* UIP, NSIP, cryptogenic organising pneumonitis, *etc.*), with some pedigrees demonstrating several subtypes of IIP within the same family. The findings from this study [16] confirm previous reports of disease heterogeneity in familial IPF within individual families [17], suggesting that the background of specific individuals affects the phenotype of disease induced by a common central mechanism [15].

Genome-wide linkage analysis in familial IPF has demonstrated linkages to a number of chromosomes in familial IPF [18, 19]. This number is increasing as more families are analysed and more data become available, although in many cases the Lod scores for these linkages are modest. With the large volume of familial information that has been assembled through the collaborative effort of Duke University, National Jewish Hospital/University of Colorado and Vanderbilt University, it is hoped that a genome-wide search will identify specific mutations responsible for pulmonary fibrosis in these families.

SURFACTANT PROTEIN C AND FAMILIAL PULMONARY FIBROSIS

Recent reports have linked some familial cases of pulmonary fibrosis to mutations in the gene encoding surfactant protein

(SP)-C. A mutation in the gene (*SFTPC*) encoding SP-C, which resulted in deletion of exon 4 of the SP-C pro-protein and an absence of mature SP-C in the bronchoalveolar lavage, was found to be associated with chronic lung disease in an infant with a family history of interstitial lung disease [20]. Using a candidate gene approach, the study of a relative with familial pulmonary fibrosis identified a heterozygous exon 5 + T128A transversion of *SFTPC*, which was thought to result in impaired processing of SP-C precursor protein (proSP-C) [17]. In this family, the adults suffered from UIP and the children had cellular NSIP. Immunostaining for proSP-C demonstrated abnormal cytoplasmic distribution in type-II epithelial cells, with diffuse cytoplasm staining in affected individuals rather than localised, dense staining associated with lamellar bodies in normal tissue (fig. 3). When these cells were examined using electron microscopy, affected individuals showed excessive numbers of abnormal structures, which appear to be defective lamellar bodies [17]. These findings suggest that, at least in this family, the origin of their lung disease is related to SP-C misfolding. Further data regarding the potential role of SP-C in the development of pulmonary fibrosis were obtained from a model in which murine lung epithelial cells transfected with a specific mutation in the endoplasmic reticulum-lumenal (C-terminal) part of proSP-C, a mutation known to be associated with lung disease, show increased cytotoxicity and apoptosis compared with normal controls [21]. To summarise, mutations associated with pulmonary disease have been identified in *SFTPC*. However, these mutations are rare, accounting for only two out of the 200 families studied by the present authors. Although *SFTPC* mutations are not the causative factors in the majority of IPF cases, these findings highlight the importance of the type-II alveolar epithelial cells in disease pathogenesis and suggest that further examination of this cell population may provide evidence regarding the underlying pathogenesis of IPF.

TELOMERES AND FAMILIAL IPF

A spectrum of different genetic diseases is associated with pulmonary fibrosis, including neurofibromatosis, tuberous sclerosis, dyskeratosis congenita and metabolic disorders, such as Niemann–Pick and Gaucher disease. Dyskeratosis congenita is a rare, progressive bone marrow failure syndrome related to telomerase dysfunction, which is caused by mutations in the genes that encode proteins in the telomerase complex responsible for maintaining telomeres at the ends of chromosomes [22]. The telomere is a region of highly repetitive DNA sequences bound by specific telomere binding proteins, which are important in protecting the end of the chromosome during replication by preventing DNA repair enzymes from recognising them as breaks in the DNA. However, if left uncovered, telomeres are subject to DNA repair which can lead to chromosomal fusion or translocation.

Telomeres are extended by the ribonucleoprotein enzyme, telomerase. Telomerase has two key components: 1) telomerase RNA (hTR or TERC), which is associated with the protein dyskerin; and 2) a specialised reverse transcriptase enzyme known as hTERT (Telomerase Reverse Transcriptases). hTERT is a very important enzyme for cell replication and maintenance of the telomere. In the X-linked recessive form of dystkeratosis congenita, the gene defect lies in the *DKC1* gene, which encodes for the protein dyskerin [23]. In the autosomal
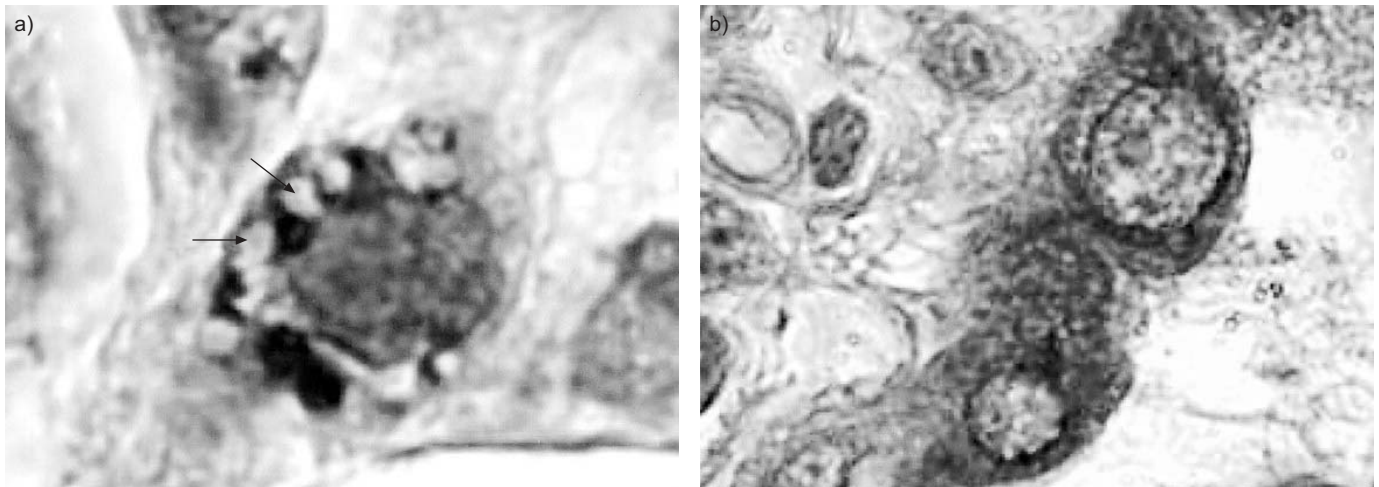
FIGURE 3. Photomicrographs of immunohistochemistry. a) Lung section of a normal subject immunostained for pro-surfactant protein (proSP)-C. Type-II cell shows predominately focal dark staining of the cytoplasm adjacent to lamellar bodies, which are evident as clear vesicles (arrows). b) Lung section immunostained for proSP-C, taken from an explanted lung from a familial pulmonary fibrosis patient. Two cuboidal type-II cells show diffuse dark cytoplasmic staining. No obvious lamellar bodies are seen. Reproduced from [17] with permission from the publisher.

dominant form, however, mutations in *hTERC* are responsible for disease phenotype [24]. Recently, evidence for a link between telomere function and IPF has come from work by ARMANIOS *et al.* [25], who identified a family with autosomal dominant dyskeratosis congenita who carried a null *hTERT* allele but lacked the typical mucocutaneous features of dyskeratosis congenita. Pulmonary fibrosis was dominantly transmitted in this family, and in one individual it was the only presenting feature.

As the clinical presentation and pattern of fibrosis in this subject were typical of the idiopathic form of the disease, and since familial IPF is also dominantly inherited, ARMANIOS *et al.* [25] developed the hypothesis that, in general, telomere shortening may be a cause of familial IPF, and that mutations in telomerase may contribute to it. To test this hypothesis, 73 pro-bands from the Vanderbilt Familial Pulmonary Fibrosis Registry (Nashville) were screened for mutations in *hTERT* and *hTR* [26]. Six (8%) pro-bands were identified with heterozygous mutations in *hTERT* or *hTR*; five of the six pro-bands did not have any of the classic features of dyskeratosis congenita. Mutant telomerase was associated with short telomeres relative to the expected median for age. Short telomeres were found in carriers of the mutation both with symptoms of IPF and also in asymptomatic subjects, suggesting that such individuals may be at risk for the disease [26].

TSAKIRI *et al.* [27] used linkage to map a gene on chromosome 5 which was associated with familial IPF in two families. Sequencing revealed multiple mutations in *hTERT* that co-segregated with pulmonary disease. Analysis of the pro-bands of 44 additional unrelated families and 44 sporadic cases of interstitial lung disease revealed five other mutations in *h*TERT and one family with a heterozygous mutation in TERC. Heterozygous carriers of all of the mutations in *h*TERT or TERC in the study by TSAKIRI *et al.* [27] also had shorter telomeres than age-matched family members without the mutations, whether or not they had symptomatic IPF (fig. 4).

These studies support a role for mutation in the telomerase genes in a proportion, perhaps 10%, of patients with familial IPF. This highlights the fact that different genes are responsible for the development of, or predisposition to, pulmonary fibrosis across families, although transmission patterns suggest that a single dominant gene will be found to be responsible within each specific family.

CONCLUSION

It is clear that the pathogenesis of IPF is complex, and that there is unlikely to be a single gene which is responsible for the majority of cases. Global expression studies in IPF have implicated a wide range of target genes, but the sheer number of candidates can make it difficult to focus on which genes are the most important
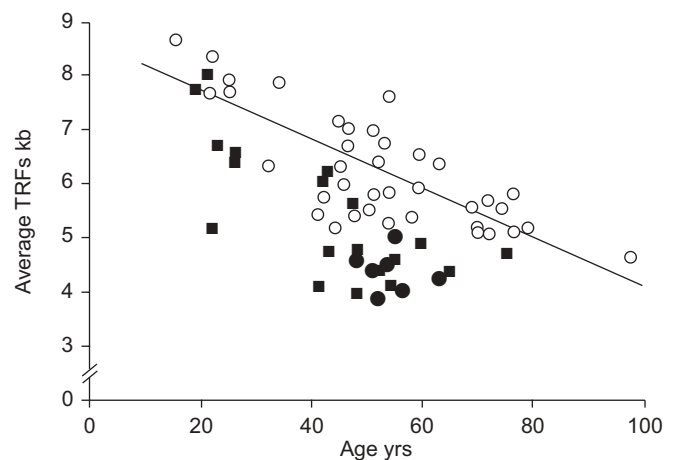

FIGURE 4. Telomere length determined by Southern blotting of chromosomal terminal restriction fragments (TRFs) of genomic DNA isolated from leukocytes. Average TRFs for each individual in all families are plotted against age. ○: normal; ■: mutation without idiopathic pulmonary fibrosis; ●; mutation with idiopathic pulmonary fibrosis. Reproduced from [27] with permission from the publisher.

and which to subject to further investigation and validation. Furthermore, studies of gene expression in a particular tissue raise numerous questions including the relevance of any changes to specific cell type, whether such changes are involved in the pathological process itself, and whether and how such genetic abnormalities correlate with protein levels.

To date, the only clearly defined genetic mutations associated with idiopathic pulmonary fibrosis are those in the gene *SFTPC* and in the telomerase genes *hTERT* and *TERC*, which have been demonstrated in certain families with familial idiopathic pulmonary fibrosis. Although rare, such mutations are beginning to provide important information regarding disease mechanisms. For example, both the surfactant protein-C and telomerase mutations are highly expressed in type-II alveolar epithelial cells, supporting a central role for these cells in the underlying mechanism of pulmonary fibrosis. Despite limitations, gene expression profiling has the potential to improve the current understanding of the underlying pathogenesis of IPF and to help identify future therapeutic targets for the management of this devastating disease.

REFERENCES

1 Costabel U, King TE. International consensus statement on idiopathic pulmonary fibrosis. *Eur Respir J* 2001; 17: 163–167.

2 Kaminski N, Allard JD, Pittet JF, *et al.* Global analysis of gene expression in pulmonary fibrosis reveals distinct programs regulating lung inflammation and fibrosis. *Proc Natl Acad Sci USA* 2000; 97: 1778–1783.

3 Zuo F, Kaminski N, Eugui E, *et al.* Gene expression analysis reveals matrilysin as a key regulator of pulmonary fibrosis in mice and humans. *Proc Natl Acad Sci USA* 2002; 99: 6292–6297.

4 Cojocaru GS, Rechavi G, Kaminski N. The use of microarrays in medicine. *Isr Med Assoc J* 2001; 3: 292–296.

5 Van Berkum NL, Holstege FC. DNA microarrays: raising the profile. *Curr Opin Biotechnol* 2001; 12: 48–52.

6 Durinck S. Pre-processing of microarray data and analysis of differential expression. *Methods Mol Biol* 2008; 452: 89–110.

7 Ben-Dor A, Bruhn L, Friedman N, Nachman I, Schummer M, Yakhini Z. Tissue classification with gene expression profiles. *J Comput Biol* 2000; 7: 559–583.

8 Pardo A, Gibson K, Cisneros J, *et al.* Up-regulation and profibrotic role of osteopontin in human idiopathic pulmonary fibrosis. *PLoS Med* 2005; 2: e251.

9 Selman M, Pardo A, Barrera L, *et al.* Gene expression profiles distinguish idiopathic pulmonary fibrosis from hypersensitivity pneumonitis. *Am J Respir Crit Care Med* 2006; 173: 188–198.

10 Yang IV, Burch LH, Steele MP, *et al.* Gene expression profiling of familial and sporadic interstitial pneumonia. *Am J Respir Crit Care Med* 2007; 175: 45–54.

11 Rosas IO, Kaminski N. When it comes to genes – IPF or NSIP, familial or sporadic – they're all the same. *Am J Respir Crit Care Med* 2007; 175: 5–6.

12 Lee HL, Ryu JH, Wittmer MH, *et al.* Familial idiopathic pulmonary fibrosis: clinical features and outcome. *Chest* 2005; 127: 2034–2041.

13 Marshall RP, Puddicombe A, Cookson WO, Laurent GJ. Adult familial cryptogenic fibrosing alveolitis in the United Kingdom. *Thorax* 2000; 55: 143–146.

14 Allam JS, Limper AH. Idiopathic pulmonary fibrosis: is it a familial disease? *Curr Opin Pulm Med* 2006; 12: 312–317.

15 Lawson WE, Loyd JE. The genetic approach in pulmonary fibrosis: can it provide clues to this complex disease? *Proc Am Thorac Soc* 2006; 3: 345–349.

16 Steele MP, Speer MC, Loyd JE, *et al.* Clinical and pathologic features of familial interstitial pneumonia. *Am J Respir Crit Care Med* 2005; 172: 1146–1152.

17 Thomas AQ, Lane K, Phillips J 3rd, *et al.* Heterozygosity for a surfactant protein C gene mutation associated with usual interstitial pneumonitis and cellular nonspecific interstitial pneumonitis in one kindred. *Am J Respir Crit Care Med* 2002; 165: 1322–1328.

18 Hodgson U, Pulkkinen V, Dixon M, *et al.* ELMOD2 is a candidate gene for familial idiopathic pulmonary fibrosis. *Am J Hum Genet* 2006; 79: 149–154.

19 Steele MP, Brown KK. Genetic predisposition to respiratory diseases: infiltrative lung diseases. *Respiration* 2007; 74: 601–608.

20 Nogee LM, Dunbar AE 3rd, Wert SE, Askin F, Hamvas A, Whitsett JA. A mutation in the surfactant protein C gene associated with familial interstitial lung disease. *N Engl J Med* 2001; 344: 573–579.

21 Mulugeta S, Maguire JA, Newitt JL, Russo SJ, Kotorashvili A, Beers MF. Misfolded BRICHOS SP-C mutant proteins induce apoptosis *via* caspase-4- and cytochrome c-related mechanisms. *Am J Physiol Lung Cell Mol Physiol* 2007; 293: L720–L729.

22 Walne AJ, Marrone A, Dokal I. Dyskeratosis congenita: a disorder of defective telomere maintenance? *Int J Hematol* 2005; 82: 184–189.

23 Heiss NS, Knight SW, Vulliamy TJ, *et al.* X-linked dyskeratosis congenita is caused by mutations in a highly conserved gene with putative nucleolar functions. *Nat Genet* 1998; 19: 32–38.

24 Vulliamy T, Marrone A, Goldman F, *et al.* The RNA component of telomerase is mutated in autosomal dominant dyskeratosis congenita. *Nature* 2001; 413: 432–435.

25 Armanios M, Chen JL, Chang YP, *et al.* Haploinsufficiency of telomerase reverse transcriptase leads to anticipation in autosomal dominant dyskeratosis congenita. *Proc Natl Acad Sci USA* 2005; 102: 15960–15964.

26 Armanios MY, Chen JJ, Cogan JD, *et al.* Telomerase mutations in families with idiopathic pulmonary fibrosis. *N Engl J Med* 2007; 356: 1317–1326.

27 Tsakiri KD, Cronkhite JT, Kuan PJ, *et al.* Adult-onset pulmonary fibrosis caused by mutations in telomerase. *Proc Natl Acad Sci USA* 2007; 104: 7552–1557.