TITLE: The derepression of transposable elements in lung cells is associated with the inflammatory response and gene activation in idiopathic pulmonary fibrosis

Mahboubeh R. Rostami[1] and Martina Bradic[2]*

[1]Department of Genetic Medicine, Weill Cornell Medical College, New York, NY
[2]Marie-Josee and Henry R. Kravis Center for Molecular Oncology, Memorial Sloan Kettering Cancer Center, New York, New York.

*corresponding author:
Martina Bradic, bradicm@mskcc.org; mb3188@gmail.com

Abstract

Background

Transposable elements (TEs) are repetitive sequences of viral origin that compose almost half of the human genome. These elements are tightly controlled within cells, and if activated, they can cause changes in both gene regulation and immune viral responses that have been associated with several chronic inflammatory diseases in humans. As oxidants are potent activators of TEs, and because oxidative injury is a major risk factor in relation to idiopathic pulmonary fibrosis (IPF), we hypothesized that TEs might be involved in the regulation of gene expression and so contribute to inflammation in cases of IPF. IPF is a fatal lung disease that involves the gradual replacement of the alveolar tissue with fibrotic scars as well as the accumulation of inflammatory cells in the lower respiratory tract. Although IPF is known to occur as a result of the complex interaction between environmental risk factors (i.e., oxidative stress) and genetics, the relative contributions of these factors to the disease remain unclear. To determine whether TEs are associated with IPF, we compared the transcriptional profiles of the genes and TEs of lung cells obtained from both healthy donors and IPF patients.

Results

We quantified the TEs and gene expression levels using a published bulk RNA-seq dataset concerning 24 subjects (16 donors and eight IPF patients), including three lung-cell types per subject, as well as a scRNA-seq dataset concerning 16 subjects (eight donors and eight IPF patients).
We found evidence of TE dysregulation in the alveolar type II lung cells and alveolar macrophages of the IPF patients. In addition, the activation of the LINE family of elements in IPF is associated with the increased expression of TE cellular regulators (MOV10, IFI16, SAMHD1, and APOBECG3), interferon-stimulating genes (ISG15, IFI6, IFI27, IFI44, and OAS1), chemokines (CX3CL1 and CXCL9), and interleukins (IL15RA). We also demonstrated that TE derepression might be involved in the regulation of previously reported IPF candidate genes (MUC5B, CHL1, CCL22, and MMP7).

Conclusion

Based on our findings, we propose that TEs derepression play an important role in the regulation of gene expression and can also prompt both the recruitment of inflammatory processes and the disruption of the immunological balance, which can lead to chronic inflammation in IPF.

Background

Over half the human genome is composed of repetitive sequences known as transposable elements (TEs). These repeated regions of the genome are organized into DNA transposons, which propagate via a cut and paste mechanism, and retrotransposons, which move by means of a copy and paste mechanism [1, 2]. Retrotransposons are the most prevalent TEs in humans, and they are further divided into the long terminal repeats (LTR) superfamily, which includes endogenous retroviruses (ERV), and the non-LTR superfamily, which includes both short interspersed elements (SINEs) and long interspersed elements (LINEs). Each of these superfamilies harbors a diverse family of repetitive sequences, and only the non-LTR elements from the L1, Alu, and SVA families can still transpose in the human genome [3]. TE transposition can destabilize the genome in many different ways, including gene disruption, the modulation of gene transcription, and mRNA processing through numerous mechanisms [4, 5]. Given their viral origin, the overexpression of TEs can also mimic viral infection and so trigger an innate immune response, leading to chronic inflammation. Although the TE activity is tightly regulated in somatic cells at the transcriptional and post-transcriptional levels, dysregulation can occur due to changes in methylation, histone modification, or mutations in the genes involved in TE regulation [6-8]. Notably, elevated cytokine levels and chronic inflammation in response to increased TE expression have previously been demonstrated in relation to various human diseases, including multiple sclerosis, systemic lupus, lateral sclerosis, Rett syndrome, Aicardi-Goutières syndrome, aging-related pathologies and complex lung disorders [9, 10]. However, due to their repetitive nature, TEs are often excluded from analysis, which explains why their effect on expression and their involvement in the processes of diseases have not been systematically studied. Here, we aim to survey the TE activity in idiopathic pulmonary fibrosis (IPF) as well as to determine whether TE

activation might be involved in both the IPF-related inflammatory response and the regulation of IPF-related genes.

IPF is a non-treatable inflammatory lung disease that involves the gradual replacement of the alveolar tissue with fibrotic scars as well as the accumulation of inflammatory cells in the lower respiratory tract [11]. The detection of IPF-associated genetic variants has enhanced our understanding of the role played by inherited risk factors in the disease risk. However, the underlying causes of IPF are not yet well understood, while vital questions persist regarding the ways in which the complex interaction between risk factors (e.g., smoking, viral infection, oxidative stress, age) and genetics causes the pathogenesis of IPF [12]. An important feature of that pathogenesis is the shift in epithelial cell populations whereby type I alveolar (AT1) epithelial cells are damaged and the epithelial surface is populated by type II alveolar (AT2) epithelial cells and bronchiolar epithelial cells. The lower respiratory tract of an IPF patient is primarily populated by alveolar macrophages (AMs) and neutrophils, which are among the first responders to cellular defense and which play a significant role in absorbing harmful particles that have passed through the mechanical barrier of the respiratory system. When activated, AMs spontaneously release toxic oxidants (i.e., H_2O_2), which place a persistent oxidative burden on the fragile structure of the alveoli, and therefore, represent one of the most important mechanisms of AT2 epithelial cell injury in cases of IPF [13, 14].

As oxidants are potent activators of TEs [15], we hypothesized that exposure to oxidative stress fosters a permissive environment in lung cells that unleashes TEs, which modify the adjacent gene expression and also contribute to chronic inflammation due to their “viral mimicry” potential. To test this hypothesis, we used published transcriptome profiles of lung cells obtained from both healthy donors and pulmonary fibrosis patients [16]. This data were chosen because it allows for the profiling of the TE activity in the individual cell populations within those cells that are highly relevant to the disease. We determined the upregulation of TEs in IPF patients using the bulk RNA sequencing data in AT2, AM, and whole-lung cells, and we confirmed the upregulated TE activity of the L1 TE family in individual cell clusters using single-cell RNA sequencing (scRNA-seq) data. This is the first study to survey and relate TE activity to IPF. Additionally, it demonstrates that TE derepression might be

involved in the regulation of the previously reported IPF candidate genes, and further, that active TEs might be involved in the perpetual inflammation of the lower respiratory tract in cases of IPF.

Results

Increased TE expression is positively correlated with both the activation of cellular TE inhibitors and the innate immune response and negatively correlated with autophagy in IPF

To determine whether the expression of TEs changes in IPF, we quantified the TE expression from previously published reports of the bulk RNA sequencing of 14 donor lung biopsies and compared it to explants from eight transplant recipients [16] (**Supplemental Table 1**). We first compared the gene and expression of the TE families between IPF patients and donors in flow cytometry-sorted AMs, AT2 cells, and whole lungs (WLs). The TE families are defined as groups of TEs with similar sequences across the genome (subfamilies), and thus, their expression is averaged per the number of such groups.

The largest changes in the expression of the TE family were identified in the AT2 cells (72 up, 22 down), with the largest number of changes being present in the LTR-TEs (**Figure 1, Table 1**). The WL that represents a mixture of different cell types exhibited 18 TE subfamily differences (14 up, 4 down), with the largest number of changes again being present in the LTR-TEs. Finally, the TE activity in the AMs was characterized by 17 TE subfamily differences (11 up, 6 down), again primarily in the LTRs. A few other TE expression changes representing DNA transposons, which are less abundant TE families in humans, were also identified. The changes in the gene expression showed the same trend as the TE changes, and they were also the highest in the AT2 cells (4131), followed by the WL (1170) and AMs (1033) (**Table 1**).

TEs are tightly controlled by multiple mechanisms, including TE promoter methylation, and the inhibitory host factors involved in transcriptional and post-transcriptional TEs control [17]. Modifications in these control points might activate TEs and cause pathological states. Therefore, we assessed whether the numerous TE expression changes that occur in AT2 cells, which are known to be a

primary target of injury in IPF [18], relate to transcriptional changes in those cellular factors known to be involved in retrotransposon activity [17, 19].

In particular, we focus on the LINE family of elements, which are only upregulated in AT2 cells (**Table 1**). These elements are responsible for the majority of retrotransposition activity in human diseases [9, 20] and their regulation has been well documented in the literature [17, 58] (**Supplemental Table 2A**). We detected differential expression of 126 out of 480 tested TE- related genes (**Supplemental Table 2B**). We also detected the significant upregulation of the Forkhead Box A1 (FOXA1) transcription factor (TF), which was found to be positively correlated with the expression of LINE L1 TEs in the IPF patients (**Figure 2, Supplemental Table 2B** and **2C**). Among its various functions, FOXA1 is known to promote L1 expression by binding to its promoter [19], which suggests the potential demethylation of the L1 promoters in IPF. In addition, we also identified the upregulation of multiple L1 inhibitory factors. For example, the sterile alpha-motif (SAM) and histidine-aspartate (HD) domain-containing protein 1 (SAMHD1), which actively inhibits retroviral replication by means of dNTP degradation [21], was upregulated, while its expression was positively correlated with multiple upregulated L1 TEs in IPF (**Figure 2 Supplemental Table 2B** and **2C**). Furthermore, the inhibitory host factors involved in post-translational control, including MOV10 and APOBEC3G, were also found to be upregulated in the IPF patients. Additionally, they were determined to be correlated with L1 expression, which indicates their activity against high L1 expression in the cells (**Figure 2, Supplemental Table 2B** and **2C**). Finally, we found that CALCOCO2 (NDP52) and SQSTM (p62), which play an important role in degrading the retrotransposon RNA in cytoplasm [22], were downregulated in the IPF patients and also negatively correlated with the expression of the L1s (**Figure 2, Supplemental Table 2B** and **2C**). This observation is suggestive of the failure of the autophagy pathway to degrade the retrotransposon RNA, which can result in the accumulation of L1 in the cells. Uncontrolled L1 accumulation results in the formation of DNA, RNA, or DNA-RNA intermediates that can activate the antiviral response, including the activation of the cytosolic sensors DDX58 (RIG-I) or IFI16, which recognize ssDNA [17]. Both these sensors were upregulated in the IPF patients as well as positively correlated with the expression of the numerous L1s. These sensors are also responsible for the downstream activation of

various inflammatory cytokines, including interferons (IFNs), chemokines, and interleukins, which play an important role in the activation of the immune response and the promotion of disease [9]. In accordance with these findings, we also detected the upregulation of those genes involved in the interferon pathway (i.e., STAT1 and IRF9), interferon-stimulating genes (i.e., ISG15, IFI6, IFI27, IFI44, and OAS1), interleukins (IL15RA), and chemokines (CX3CL1and CXCL9), which were all positively correlated with the TE expression in the IPF patients (**Figure 2, Supplemental Table 2B** and **2C**). The difference between the donors and the IPFs was also significant for most of these genes when comparing correlation slopes determining relationship among L1 subfamily upregulation and TE-related genes in AT2 cells (**Supplemental Table 2D**).

Locus-level TE expression is associated with IPF and is also cell-type specific

The induction of the immune response by means of the upregulation of multiple TE families does not constitute the only detrimental effect of TEs. Indeed, TEs can also trigger changes in gene expression at the locus level. To understand the impact of IPF-related TE activity on nearby genes, we also quantified the changes in TE expression that occurred at individual genomic locations within the three cell types. We again identified the largest number of changes in the AT2 cells (1489 up, 1149 down), while the WL also featured an abundant number of changes representing 1341 TE loci (588 up, 753 down) (**Figure 3, Supplemental Table 3**). Finally, the TE activity in the AM was found to be characterized by 359 TE loci (193 up, 166 down). We first intersected the upregulated TE locus changes as well as the gene changes between the IPF patients and the healthy donors in each cell type (**Figure 4A**). This analysis revealed that most changes were unique and only present in an individual cell type (AT2 = 1426, AM = 180, and WT = 530), although 53 TE changes were shared between the AT2 and WL, four between the WL and AM, and nine between the AM and AT2, while one change was common to all three comparisons. This suggests the potentially strong impact of TE activity specific to the AT2 and AM that might affect the gene expression related to IPF and that is not detected when analyzing the heterogeneous WL cell population. The number of gene expression

changes shared among the different cell types also showed a similar trend, with the highest number of expressed genes being observed in the AT2 (3826), followed by the AM (822), and the WT (644).

We next examined the distribution of the TE changes across the genomic regions. To do so, all the differentially expressed TE changes were categorized as belonging to either the 5'UTR, 3'UTR, exon, intron, or intergenic region. These analyses revealed the same pattern of changes as observed in relation to the three comparisons (**AT2, AM, and WL, Figure 4B, Supplemental Table 3**), with the highest number of TE transcriptional changes being seen in the 3'UTR region, while the smallest number of changes were found in the intron and intergenic regions. The changes in the exon region as well as the exon-intron, exon-5'UTR, and exon-3'UTR changes were excluded from further analysis because they were most likely related to transcriptional noise.

We also examined the intergenic TE loci with the highest differential expression in the AT2 cells between the IPF patients and the donors in order to determine whether they were part of the intrinsically active elements or whether their expression extended beyond the boundaries of the TEs. We present an example of the intergenic L1PA3 element, which was located ~32.3 kb away from the closest gene (**Figure 5A**) suggesting its intrinsic regulation. Another example of the intergenic TEs is shown ~5.3 kb from the long non-coding RNA (AC023137.1), in which the L1PA7 differs in terms of its expression between the IPF patients and the donors, and further, its expression does not extend the same coverage beyond the boundaries of the element (**Figure 5B**).

TE expression is correlated with the gene expression of several IPF candidate genes

To further illustrate the importance of locus-specific TEs, we tested the effects of the intragenic and intergenic upregulated TE (excluding those in the exons) loci on the transcriptional regulation of the genes. As the pairwise comparisons of each gene with each TE locus identified a large number of changes, some of which proved difficult to interpret, we focused on the TEs matched with their adjacent genes (cis interactions) and examined whether TE expression predicts nearby gene

expression. There were 1512 pairs (out of 10092988 tests conducted between 3826 genes and 2638 TEs, with 7.80% being significant pairs) in the AT2 cells, 198 pairs in the AM cells (out of 295098 tests conducted between 822 genes and 359 TEs, with 1.4% being significant pairs), and 497 pairs in the WL (out of 863604 tests conducted between 644 genes and 1341 TEs, with 0.04% being significant pairs) that matched the cis criteria (**Figure 6**). We identified a total of 172, 105, and four significant gene-TE pairs ($p < 0.05$) from among our cis subsets in the AT2, AM2, and WL, respectively. In some cases, there were more than one TE associated with the expression of the same gene. Thus, we further classified the difference into the number of unique genes whose transcription is significantly related to the TEs. This classification resulted in the identification of a total of 107 genes in the AT2 cells, 63 genes in the AM cells, and three genes in the WL tissue that had a transcriptional pattern correlated with TEs and that differed between the IPF patients and the healthy donor (**Figure 6**). Interestingly, although a high number of TE-gene pairs were present in the WL comparison, only a very few significant differences were observed, which again suggests that important signatures might be missed in heterogeneous cell populations. Our gene ontology enrichment analysis of the genes found to be correlated with TE expression in the AT2 cells identified cilium movement as well as axoneme and organelle assembly related processes (FDR < 0.005) (**Table 2A**). This suggests that the activation of TEs can result in the activation of genes that might be involved in cellular identity changes. In the AM cells, the TE-associated gene expression was found to be linked with immune-related cellular process.

Most of the TEs found to be associated with gene regulation in the AT2 cells were identified in the introns (75), followed by the 3'UTR region (82), outside the gene (12), and the 5'UTR region (three) (**Supplemental Table 4A, Table 2B**). The SINE (Alu) elements were most commonly found to be associated with gene expression in the AT2 cells, and they were mostly embedded within the 3'UTR region of the genes. Several of the genes found to be significantly correlated with TEs represent important IPF candidate genes that have been identified in multiple genome-wide association studies or functional studies related to IPF [23]. For example, we identified a 3.88-fold expression change

between the IPF patients and the healthy donors in the TE (chr11|1253519|1253937|MLT1C:ERVL-MaLR:LTR) that is located between exons 33 and 34 and that is significantly correlated with the high expression of the MUC5B gene in IPF (**Supplemental Table 4A**). Furthermore, the relationship between the gene and the TE expression of IPF group is significantly different than the same relationship in donor group (**Figure 7A,** p.adjust=0.0284). This suggests the potential role played by TEs in the regulation of this gene, which in turn plays an important role in mucin excretion and significantly contributes to IPF pathogenesis [24, 25]. Notably, we identified the association between one of the three core IPF gene markers, namely the cell adhesion molecule L1-like (CHL1) gene, and the L1PA6 TE (chr3|367661|374053|L1PA6:L1:LINE) (**Supplemental Table 4A**). However, two other TEs (chr3|391232|391535|AluSz:Alu:SINE, chr3|408347|408547|MIRc:MIR:SINE|342), also showed a significant association with CHL1 expression. To determine which of the TEs made the largest contribution to CHL1 gene expression, we used all three TEs in a linear model and then calculated the relative importance of each of them in terms of influencing CHL1 expression. The proportion of the variance explained by the model containing all three TEs was 86.09%. Further, each TE explained ~20% of the variance, while the group covariate representing the phenotype explained ~25% of the variance, thereby suggesting the similar contributions of all the TEs to the gene expression. Significant difference between the correlation slope in IPF and donor groups was also observed for L1PA6- CHL1 relationship (**Figure 7A**, p.adjust= 0.0019) and the other two TEs (AluSz, MIRc) associated with CHL1. This further confirms regulation of CHL1 by adjacent TEs in IPFs.

The serpin family B member 3 (SERPINB3) is another important IPF candidate gene for which we identified a 4.23-fold change in the gene expression between the IPF patients and the healthy donors, which was correlated with a 4.21-fold change in the LINE TE element (chr18|63651629|63653187|L1MA8:L1:LINE) (**Supplemental Table 4A**). This element is located ~2 kb downstream of the gene. Glutathione S-transferase alpha 2 (GSTA2) also exhibited upregulation in the IPF patients, and its expression was associated with the LTR5A TE (chr6|52748278|52749305|LTR5A:ERVK:LTR) located 783 bp downstream of that gene (**Figure 7A**, **Supplemental Table 4A**). Although TEs significantly predicts the expression of SERPINB3 and

GSTA2 genes, that relationship does not significantly differ between IPF and donors as shown by test for correlation slopes (**Figure 7A**, p.adjust=0.1260, p.adjust=0.9227, respectively). Nevertheless, significant difference of TEs expression results in significant difference in gene expression for these genes in IPF suggesting their role in the disease.

The analysis of the TE loci in the AM cells primarily determined the association between gene expression and the SINE elements (56), followed by the LINE (26), LTR (six), and 17 other elements (e.g., DNA transposons) (**Supplemental Table 4B, Table 2B**). Most of the elements from all the families were activated within the 3'UTR region, and we did not identify any elements associated with gene expression outside the gene. We found that the expression of the chemokine CCL22 was related to the three SINE elements **(**chr16|57364478|57364778|AluSq2:Alu:SINE, chr16|57364806|57365101|AluSz:Alu:SINE, chr16|57365617|57365740|AluJo:Alu:SINE) located in the 5'UTR and 3'UTR region (**Table 4B**). This chemokine contributes to activation of alveolar macrophages and subsequently to the lung damage in patients with IPF [26]. The proportion of variance explained by the model predicting the CCL22 expression was 93.71%, with AluSz (3'UTR) explaining ~29% and AluJo (3'UTR) and AluSq2 (5'UTR) explaining ~27% and ~24%, respectively. Correlation slopes for IPF and donors are also significantly different for two of these elements providing additional evidences for strong regulation of CCL2 by TEs activation in IPF patients (**Figure 7B,** only plot for AluSq2 element is shown, p.adjust= 0.0237**).** We also found that the upregulated TE expression in the intron (chr14|22843913|22844208|AluSx1:Alu:SINE), and in the 3'UTR region (chr11|102520549|102520704|L2c:L2:LINE) significantly correlates with two upregulated matrix metalloproteases (MMP14 and MMP7, respectively) in IPF group (**Supplemental Table 4B**). MMPs are important players in cell migration and tissue repair in lungs and have been related to IPF pathogenesis [27]. Here, we establish correlation of two MMPs with TEs and we further identify significant TE-gene correlation slope difference in MMP7 between IPF and donors which further confirms differential regulation of MMP7 in IPF **(Figure 7B,** p.adjust= 0.0156).

We were also able to associate a number of other genes with TE expression that have previously been related to IPF, including osteopontin (SPP1) in the 3'UTR region (Chr4: 87979662-87979893) and the interleukin 1 receptor antagonist (IL1RN) within the 5'UTR region (chr2|113133292|113133561|Charlie18a:hAT-Charlie:DNA) (**Supplemental Table 4B**). Both these genes are known to be expressed at high levels in IPF AM cells, while previously reported immunohistochemistry results have confirmed that these markers are not expressed in donor tissue [16]. Our study also finds significant upregulation of these genes and their adjacent TE loci in IPF. However regression slopes between donors and IPF did not differ for these two TE-gene pairs, suggesting that TEs might be involved in the regulation of these genes in health and in the disease (Figure 7B, p.adjust= 0.1739 for SPP1, p.adjust= 0.8381 for IL1RN) .

ScRNA-seq analysis confirms the TE changes in multiple cell populations in the fibrotic human lung

We also analyzed a previously published dataset concerning eight donors and eight IPF patients that had been generated by means of scRNA-seq technology [28]. A total of 77,517 single cells and 22,009 genes were obtained. We assigned each cluster to a cell type based on the expression of the established markers in that cluster (**Supplemental Figure 1A**), and the following cell types were confirmed: epithelial cells (alveolar type II cells (AT2), alveolar type I cells (AT1), ciliated cells, basal cells, and club cells), immune cells (alveolar macrophages (AMs), monocytes, B cells, plasma mast cells, dendritic cells, and T cells), and mesenchymal cells (fibroblasts and endothelial cells) (**Figure 8A**). The distribution and identity of the cells were similar between the two phenotypes (**Figure 8B, Supplemental Table 5A**). To confirm the differential TE activity in the AM2 and AT2 cells in the IPF patients, as well as to determine whether the TEs differed in the other cell types, we performed a differential gene expression (DGEs) analysis between the donors and the IPF patients with regard to the TE subfamilies in each cell type (**Supplemental Table 5B)**. The results of the DGEs analysis between the IPF patients and the donors for each identified TE subfamily and all the genes in the individual lung cell type are presented in **Supplemental Table 5C** and summarized in **Figure 8C**. We

identified the increased transcription of two TE families (LINE and LTR) in the IPF patients when compared with the donors in multiple cell types (p.adjust < 0.05 and lfc > 0). Consistent with the TE upregulation identified in our RNA bulk analysis, we identified 27 significantly upregulated L1 subfamilies and eight significantly upregulated LTR subfamilies in the single-cell transcriptome. Similarly, 22 LINE subfamilies and 12 LTR subfamilies were found to be upregulated in the AM cells (**Figure 8C**). In addition, we also detected upregulation in the case of the IPF patients in the monocytes as well as the upregulation of a smaller number of subfamilies in the club, ciliated, B, dendritic, AT1, T, endothelial, and plasma cells.

Aside from identifying changes in multiple subfamilies, we were also interested in determining whether an association exists between the total expression of the L1 subfamilies and IPF in AT2 cells, as identified by bulk RNA-seq. Thus, we calculated the L1 score that represents the average expression of the upregulated L1 family per each cell type. The comparison between the L1 scores of the healthy donors and the IPF patients indicated that the L1 score is significantly higher among fibrosis patients in the AT2 cells, which indicates that it might be involved in IPF pathogenesis (**Figure 9A**). To further confirm whether the relationship between L1 activity (L1 score) and TE-related genes differs among the IPF patients and the healthy donors, we tested 126 genes which were identified in bulk RNA-seq. We identified 27 TE regulation and inflammatory response related genes (p.adjust<0.05) (i.e. APOBE3G, STAT1, SAMHD1, IRF9) in the AT2 cells (**Supplemental Table 5D**, **Figure 9B**). This observation indicates that the L1 upregulation, also confirmed in independent dataset by scRNA-seq might promote the inflammatory response in the AT2 cells of IPF patients. Single-cell RNA sequencing is a powerful method but only generates short reads from one end of a cDNA template, limiting the mapping of highly similar TE sequences. Thus, loecus specific TE expression was not performed for scRNA seq dataset.

Discussion

TEs and inflammation in IPF

The derepression of TEs can cause changes at the transcriptional and post-translational levels that involve gene expression changes, and further, that recruit immune signaling pathways that might result in pathologies [5, 29-31]. Our study highlights the involvement of TEs in IPF, and we provide evidence that TEs are activated somatically in the AT2 cells of fibrotic lungs, potentially due to injury (i.e., oxidative stress). Aside from the activation of TEs in AT2 IPF cells, we also find that the dysregulation of TEs is likely further facilitated by the decreased expression of the autophagy genes SQSTM1(p62) and CALCOCO2 (NDP52), which are known to be crucial receptors for the detection and removal of at least one TE RNA family (LINE). Dysfunctional autophagy has also been previously associated with IPF, with the suggestion being that it promotes the epithelial–mesenchymal transition of the AT2 cells contributing to fibrosis [32]. Moreover, the failure of the autophagosome removal of TEs leads to cytoplasmic TE accumulation, which can result in genome instability and inflammation [22] and, in turn, potentially in IPF.

Pulmonary fibrosis is known to be accompanied by innate and adaptive immune responses; however, the role of inflammation in the disease remains unclear [33]. We propose that TE activation and accumulation in AT2 cells might represent an important trigger of the viral cellular sensors and the activation of the innate immune system (macrophages), thereby resulting in the disruption of the immunological balance, which can cause chronic inflammation and disease. Many of the inflammation-related processes that we associate with TEs have also been previously described as contributing to IPF (IFI6, IFI27, IFI44, OAS1, IL15RA, CX3CL1, and CXCL9), and they are known to be involved in both fibroblast activation and the accumulation of the extracellular matrix [34]. Furthermore, we confirm some of these processes in our single-cell data analysis, which also shows the highest TE LINE upregulation in the AT2 cells as well as its correlation with some of the genes related to the inflammatory response in IPF patients (i.e., IFI16, IFI27, CCL2 and STAT1).

Further evidence that TEs might be involved in IPF and linked to the inflammatory response is provided by the age-related onset of IPF. Indeed, IPF is most prevalent in individuals aged 60 years or

older [35], and one of the major triggers of age-associated inflammation is the activation of the LINE TEs by the age-related loss of epigenetic marks [19, 36]. In accordance with these findings, our study also shows the significant upregulation of the FOXA1 TF that can bind to the demethylated LINE promoters and so induce LINE expression. This results in the activation of the cytosolic sensors for ssDNA (i.e., IFI16), which signals through the protein STING and induces both interferon-related changes and age-associated inflammation. Here, we also provide evidence of changes in these pathways, and we propose that some such changes might be facilitated by age-related changes.

TEs regulate genes in IPF

The inflammatory response is not the only outcome of uncontrolled global TE expression, as the activity of individual TE loci can also play an important role in modulating the expression of adjacent genes [31]. Our data show that the TE expression in the vicinity of several genes is associated with the gene expression in both the AT2 and AM cells in the IPF patients. Notably, some of these genes have previously been associated with IPF. For example, the MUC5B promoter variant is one of the major IPF risk factors associated with an increase in gel-forming mucin, which produces mucosal host defensive dysfunction in the bronchi and so is critical in IPF [37]. Our observation that TEs might be involved in the regulation of this gene is based on the significantly higher expression in IPF patients noted from TEs embedded in the intronic region of the gene. Other examples relevant to TEs include the CHL1 gene, which exhibits > 0.8 specificity and 0.9 sensitivity in distinguishing IPF patients from healthy controls, meaning that it is a potential IPF drug target [38]. Although no studies have yet related the TEs in CHL1 with IPF, previous studies have proposed that the L1P6 TE within CHL1 can act as an L1 antisense promoter and so drive the transcription of chimeric transcripts [39]. Our study also shows the upregulation of L1P6 to explain ~25% of the CHL1 expression, which indicates that this gene might potentially play a regulatory role in relation to CHL1 and so contribute to IPF. One of the fundamental genes associated with the control of proteolysis (SERPINB3) [40], together with the GSTA2 gene, which has previously been associated with IPF [41], show potential regulation from TEs outside the gene. Both these TEs are located in close downstream proximity (within 2 kb) to the gene,

and while they were not formerly described in other studies, they might have an AT2 cell-specific regulatory function. Although the majority of TE changes are identified in the AT2 cells, the AM cells also show the possible activation of metalloproteases (MMP7 and 14), interleukin (IL15RA), chemokine (CCL222), and osteopontin (SPP1), which are all implicated in the development of IPF [27]. The AM are crucial players in terms of the respiratory system defense that involves the absorption of harmful particles and infectious agents. Such processes could trigger genome instability and also activate TEs and their adjacent genes.

The TE locus expression changes in our study that are associated with the adjacent gene expression are mainly located in either the 3'UTR, intronic, or 5'UTR region of the respective genes. Previous studies have shown that TEs can be found in gene regions and impact gene regulation by acting as, or interfering with, the regulatory elements in different tissue [42]. In particular, the SINE (Alu) elements that we identify as commonly embedded and expressed in the 3'UTR region are preferentially located in gene-rich regions due to their size (300 bp). The embedded Alu sequences can regulate the translation of their host genes servin9g as cis elements, or they can be involved in the microRNA regulatory network and many other regulatory process [43]. The role of these sequences in regulating the genes within lung cells as well as their relationship with disease have yet to be determined.

TEs are sometimes retained within introns and transcribed before the transcripts are processed, or else they are sometimes not spliced out at all, as indicated by a detailed study of the LINE elements [44]. The TEs retained within introns might result from non-preprocessed RNA and so might represent transcriptional noise rather than a biological signal. Thus, it must be noted that our study has limitations related to the quantification of authentic, independently transcribed TEs. Aside from these limitations, we could detect specific signals in the AT2 and AM cells that might be relevant to IPF and that were repeated in the single-cell data analysis. In addition, although only a small number of patients were examined, our profiling of the three cell types obtained from the same patients suggests

the strongest signal to occur in the AT2 cells, which are known to be injured in IPF. This study, therefore, identifies numerous candidate loci for further functional studies.

Finally, smoking has been identified as an important risk factor for the development of IPF [45]. Interestingly, our GO term analysis of the TE loci-associated gene expression highlighted the enrichment of a few genes (eight out of 111, $p < 0.0178$) in an earlier study [46]. This earlier study compared the gene expression of the small airway epithelium (SAE) cells of smokers and non-smokers, and it found that the eight genes (i.e., GSTA2, ITGA2, KRT19, MS4A8, MUC20, MUC5B, SCGB1A1, and SNTN) seen to correlate with TE expression in our current study were also dysregulated in smokers. The SAE is the primary area where the early appearances of the majority of smoking-induced lung diseases are noted [47]. Thus, our observation might suggest that oxidative stress from cigarette smoke also represents an important TE-activating agent [48], and further, that it could be important during the very early stage of IPF development, a hypothesis that should be further tested.

Conclusion

Taken together, our findings suggest a strong link between TE expression and processes known to be key to IPF. Yet, the extent to which the dysregulation of TEs drives IPF, or whether TE activation represents a side effect of pathogenesis, remains unclear. It is, however, tempting to speculate that a combination of TE demethylation due to aging (loss of epigenetic marks) and injury (oxidative stress) accompanied by dysfunctional autophagy can lead to perpetual inflammation and to changes in locus-specific gene expression, which might play a critical role in the development of IPF in genetically predisposed individuals. In addition, our findings indicate potential new venues for therapies. Moreover, they call into question the role of TEs in other lung conditions caused by injury and inflammation (e.g., chronic obstructive pulmonary disease).

Materials and Methods

Data

A detailed description of the samples is provided in Table 1 [16]. Briefly put, we used samples from 14 donors and eight pulmonary fibrosis patients for which bulk RNA sequencing data concerning the AT2, AM, and WL cells were available. An additional eight donor and eight IPF patient WL samples were prepared and sequenced by means of 10x Chromium single-cell sequencing. All the methods used for sample processing and sequencing have previously been described (see Supplementary Table 1) [16].

Quantification of the gene expression and TE activity using bulk RNA-seq

The raw sequencing data were downloaded from the dbGAP database (phs001750.v1.p1) and then processed using the SQuIRE set of tools, which integrates the alignment and expression counts for the gene expression and TE expression [49]. We mapped the raw RNA-seq reads to the GRCh38/hg38 (Dec. 2013 release) version of the human genome assembly, which was downloaded from the UCSC Genome FTP site (ftp://hgdownload.cse.ucsc.edu/goldenPath/hg38/bigZips/chromFa.tar.gz), and determined the transcriptional changes based on both the TE family and TE locus. The counts were normalized between the samples, and the differential expression between the healthy donors and the IPF patients was determined using the *DESeq2* package in R. Significance of the TE expression was further determined using the Benjamini and Hochberg (BH)-adjusted p-value [50]. The volcano plots were constructed using the *ggplot2* function in R.

To identify the genomic locations of the differentially expressed TE loci in the cells, we downloaded the bed file annotation for the 3'UTR, 5'UTR, intron, and exon regions for each individual gene from the UCSC database table [51]. We further intersected the TE bed file coordinates of our differentially upregulated TEs with the different genomic regions using the *GenomicRanges* [52] and *regionR*

packages in R [53]. Lastly, we identified the locations of the intergenic regions (those without overlap in 3'UTR, 5'UTR, intron, and exon regions) of the differentially expressed TEs and their distance from the nearest genes using the BEDTOOLS function *closest* [54].

Association between gene expression and TE expression

We first correlated relationship between the differentially expressed TEs subfamilies and the differentially expressed TE-related genes (manually curated list of genes known to be part of the L1 defense and antiviral interferon-stimulating genes [17, 58]) using *rcorr* function in R and Pearson's correlation. The difference between the IPF and the donor groups for this relationship was further tested using F-test with *aov* function in R where TE subfamily expression represented the independent variable, and gene expression the dependent variable.

To determine whether the TE expression is related to the adjacent gene expression, we tested the correlation between the TEs located in the intergenic regions and those located within the 5'UTR, 3'UTR, and intron regions using a linear model in which the TE expression in each locus was modeled as the independent variable and the gene expression as the dependent variable, accounting for the phenotype (IPF or donor) as a factor. The TEs that overlapped with the exons or that were located within the exon-intron, exon-5'UTR, or exon 3'UTR regions were excluded from these analyses. We only tested the cis TE-gene pairs, as most of the TEs were either within 50 kb of the gene or within gene regions. We defined the significant TE-gene pairs using the multiple test correction, BH-adjusted p-value [50]. The same tests were performed for the AT2, AM, and WL data. Difference between the the regression slopes in groups (IPF or donor) was determined using F-test with *aov* function in R. Results were plotted using *ggplot2* function in R.

When the expression of more than one TE adjacent (within 50 kb distance) to the gene of interest was associated with the gene expression, we calculated the relative importance of each TE using the R

package *relaimpo* and 1000 bootstraps [55]. We further tested for the gene ontology enrichment of the genes that were associated with TE expression using the STRING database [56].

Detection of TE activity by means of scRNA-seq

Single-cell transcript mapping

The raw reads of the single-cell RNA-seq data were downloaded from dbGAP (phs001750.v1.p1) [16]. The TE annotation library was downloaded using the *SQuIRE* package in R and the GRCh38/hg38 (Dec. 2013 release) genome, together with the TE reference. The genome-TE reference was built and the reads were mapped to both the genes and TE subfamilies using Chromium 10x workflow and the Chromium 10x Cell Ranger Single-Cell Software Suite (http://software.10xgenomics.com/single-cell).

The subfamily of TEs represents groups of sequences from the TE family (LINE, SINE, and LTRs) that are sufficiently distinct in terms of their repeats to allow for unique read mapping. Only the unique alignments were considered and counted for the differential expression analysis. The number of differences is defined as the number of TE subfamilies. The cell tags were matched with the published data matrix so that the same cells were used as in previously published accounts [16]. The filtered feature matrices produced by the cell ranger were used in the subsequent analysis using the R package *Seurat* [57].

ScRNA-seq clustering

The downstream single-cell analysis was performed using functions in both the *Seurat* package V3 and R 3.5 [57]. First, we matched the cells of eight healthy donors and eight IPF patients from our alignment counts to the published filtered dataset available at GEO (GSE122960) [16]. In doing so, we recovered a total of 77,517 cells. To compare the expression data from different patients and different lung cell types, we integrated the data with the *IntegrateData* function and identified the anchors between the dataset using the *FindIntegrationAnchors* function. The normalization of the 22,009 identified genes was performed based on the total number of unique molecular identifiers (UMIs) per

cells, multiplied by a scale factor (10,000) and then log transformed. We next conducted a principal component analysis (PCA) of the top 2000 variable genes and then used the first 20 PCA components to project the cells onto a two-dimensional map using the uniform manifold approximation and projection (UMAP) dimensionally reduction method offered by the *RunUMAP* function. To identify the cell types from the lung tissue, we clustered all the IPF patient and donor cells using the K-nearest neighbor (KNN) graph-based clustering algorithm and the *FindNeighbors* function. Finally, we used the *FindClusters* function (resolution parameter = 0.5) to establish the cell clusters, while the cluster identity was determined based on the conventional cell marker genes [16, 28]. The differential gene and TE subfamilies expression analysis between the IPF patients and the donors was performed for each cell type using the non-parametric Wilcoxon rank-sum test and the Bonferroni correction. In addition, we compared the IPF patients to the donors in terms of the total expression of the L1 TE subfamilies per cell (L1 score) using the *addModuleScore* function and the Wilcoxon rank-sum test. To calculate the L1 score per cell, we used the average expression of 70 L1 subfamilies. We next tested whether the relationship between the L1 score and the TE related genes (126 genes known to be part of the L1 defense and antiviral interferon-stimulating genes identified in RNA bulk sequencing) differed between AT2 cells of the IPF patients and the donors. A linear model was fitted for each L1 score and the gene pairs for the IPF patients and the donors using the *lm* function in R, while the difference between the regression lines of two groups was determined using the t-test and *aov* function in R.

Acknowledgments

We wish to thank dbGap and Reyfman *et al.* [16] for making this data available.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

All data generated or analyzed during this study are included in Reyfman *et al.* [16] published article and its supplementary information files. Raw genomics reads and detailed patient data are available from the GSE122960 dataset and from the dbGAP database (phs001750.v1.p1), although restrictions apply with regard to the availability due to dbGAP policy.

Funding

Mahboubeh R. Rostami were supported by T32 HL094284.

Competing interests

The authors declare that they have no competing interests

Author contributions

M.B. designed the study, performed data analyses, and wrote the manuscript. M.R. performed single cell data analysis and wrote the manuscript. All authors read and approved the final manuscript.

References

1. Seberg O, Petersen G: **A unified classification system for eukaryotic transposable elements should reflect their phylogeny**. *Nature Reviews Genetics* 2009, **10**(4):276-276.
2. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W *et al*: **Initial sequencing and analysis of the human genome**. *Nature* 2001, **409**(6822):860-921.
3. Mills RE, Bennett EA, Iskow RC, Devine SE: **Which transposable elements are active in the human genome?** *Trends Genet* 2007, **23**(4):183-191.
4. Feschotte C: **Transposable elements and the evolution of regulatory networks**. *Nat Rev Genet* 2008, **9**(5):397-405.
5. Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, Imbeault M, Izsvak Z, Levin HL, Macfarlan TS *et al*: **Ten things you should know about transposable elements**. *Genome Biol* 2018, **19**(1):199.
6. Goodier JL, Kazazian HH, Jr.: **Retrotransposons revisited: the restraint and rehabilitation of parasites**. *Cell* 2008, **135**(1):23-35.
7. Baylin SB, Herman JG, Graff JR, Vertino PM, Issa JP: **Alterations in DNA methylation: a fundamental aspect of neoplasia**. *Adv Cancer Res* 1998, **72**:141-196.
8. Deniz O, Frost JM, Branco MR: **Author Correction: Regulation of transposable elements by DNA modifications**. *Nat Rev Genet* 2019, **20**(7):432.

9. Saleh A, Macia A, Muotri AR: **Transposable Elements, Inflammation, and Neurological Disease**. *Front Neurol* 2019, **10**:894.
10. Sargurupremraj M, Wjst M: **Transposable elements and their potential role in complex lung disorder**. *Respir Res* 2013, **14**:99.
11. Barratt SL, Creamer A, Hayton C, Chaudhuri N: **Idiopathic Pulmonary Fibrosis (IPF): An Overview**. *J Clin Med* 2018, **7**(8).
12. Kaur A, Mathai SK, Schwartz DA: **Genetics in Idiopathic Pulmonary Fibrosis Pathogenesis, Prognosis, and Treatment**. *Front Med (Lausanne)* 2017, **4**:154.
13. Wynn TA, Ramalingam TR: **Mechanisms of fibrosis: therapeutic translation for fibrotic disease**. *Nat Med* 2012, **18**(7):1028-1040.
14. Camelo A, Dunmore R, Sleeman MA, Clarke DL: **The epithelium in idiopathic pulmonary fibrosis: breaking the barrier**. *Front Pharmacol* 2014, **4**:173.
15. Giorgi G, Marcantonio P, Del Re B: **LINE-1 retrotransposition in human neuroblastoma cells is affected by oxidative stress**. *Cell Tissue Res* 2011, **346**(3):383-391.
16. Reyfman PA, Walter JM, Joshi N, Anekalla KR, McQuattie-Pimentel AC, Chiu S, Fernandez R, Akbarpour M, Chen CI, Ren Z *et al*: **Single-Cell Transcriptomic Analysis of Human Lung Provides Insights into the Pathobiology of Pulmonary Fibrosis**. *Am J Respir Crit Care Med* 2019, **199**(12):1517-1536.
17. Goodier JL: **Restricting retrotransposons: a review**. *Mob DNA* 2016, **7**:16.
18. Sisson TH, Mendez M, Choi K, Subbotina N, Courey A, Cunningham A, Dave A, Engelhardt JF, Liu X, White ES *et al*: **Targeted injury of type II alveolar epithelial cells induces pulmonary fibrosis**. *Am J Respir Crit Care Med* 2010, **181**(3):254-263.
19. De Cecco M, Ito T, Petrashen AP, Elias AE, Skvir NJ, Criscione SW, Caligiana A, Brocculi G, Adney EM, Boeke JD *et al*: **L1 drives IFN in senescent cells and promotes age-associated inflammation**. *Nature* 2019, **566**(7742):73-78.
20. Beck CR, Collier P, Macfarlane C, Malig M, Kidd JM, Eichler EE, Badge RM, Moran JV: **LINE-1 retrotransposition activity in human genomes**. *Cell* 2010, **141**(7):1159-1170.
21. Herrmann A, Wittmann S, Thomas D, Shepard CN, Kim B, Ferreiros N, Gramberg T: **The SAMHD1-mediated block of LINE-1 retroelements is regulated by phosphorylation**. *Mob DNA* 2018, **9**:11.
22. Guo H, Chitiprolu M, Gagnon D, Meng L, Perez-Iratxeta C, Lagace D, Gibbings D: **Autophagy supports genomic stability by degrading retrotransposon RNA**. *Nat Commun* 2014, **5**:5276.
23. Martinez FJ, Collard HR, Pardo A, Raghu G, Richeldi L, Selman M, Swigris JJ, Taniguchi H, Wells AU: **Idiopathic pulmonary fibrosis**. *Nat Rev Dis Primers* 2017, **3**:17074.
24. Seibold MA, Wise AL, Speer MC, Steele MP, Brown KK, Loyd JE, Fingerlin TE, Zhang W, Gudmundsson G, Groshong SD *et al*: **A common MUC5B promoter polymorphism and pulmonary fibrosis**. *N Engl J Med* 2011, **364**(16):1503-1512.
25. Hunninghake GM, Hatabu H, Okajima Y, Gao W, Dupuis J, Latourelle JC, Nishino M, Araki T, Zazueta OE, Kurugol S *et al*: **MUC5B promoter polymorphism and interstitial lung abnormalities**. *N Engl J Med* 2013, **368**(23):2192-2200.
26. Yogo Y, Fujishima S, Inoue T, Saito F, Shiomi T, Yamaguchi K, Ishizaka A: **Macrophage derived chemokine (CCL22), thymus and activation-regulated chemokine (CCL17), and CCR4 in idiopathic pulmonary fibrosis**. *Respir Res* 2009, **10**:80.
27. Pardo A, Cabrera S, Maldonado M, Selman M: **Role of matrix metalloproteinases in the pathogenesis of idiopathic pulmonary fibrosis**. *Respir Res* 2016, **17**:23.
28. Reyfman PA, Walter JM, Joshi N, Anekalla KR, McQuattie-Pimentel AC, Chiu S, Fernandez R, Akbarpour M, Chen CI, Ren ZY *et al*: **Single-Cell Transcriptomic Analysis of Human Lung Provides Insights into the Pathobiology of Pulmonary Fibrosis**. *Am J Resp Crit Care* 2019, **199**(12):1517-1536.
29. Sharma R, Rodic N, Burns KH, Taylor MS: **Immunodetection of Human LINE-1 Expression in Cultured Cells and Human Tissues**. *Methods Mol Biol* 2016, **1400**:261-280.

30. Payer LM, Burns KH: **Transposable elements in human genetic disease**. *Nat Rev Genet* 2019.
31. Chuong EB, Elde NC, Feschotte C: **Regulatory activities of transposable elements: from conflicts to benefits**. *Nat Rev Genet* 2017, **18**(2):71-86.
32. Hill C, Li J, Liu D, Conforti F, Brereton CJ, Yao L, Zhou Y, Alzetani A, Chee SJ, Marshall BG *et al*: **Autophagy inhibition-mediated epithelial-mesenchymal transition augments local myofibroblast differentiation in pulmonary fibrosis**. *Cell Death Dis* 2019, **10**(8):591.
33. Heukels P, Moor CC, von der Thusen JH, Wijsenbeek MS, Kool M: **Inflammation and immunity in IPF pathogenesis and treatment**. *Respir Med* 2019, **147**:79-91.
34. Agostini C, Gurrieri C: **Chemokine/cytokine cocktail in idiopathic pulmonary fibrosis**. *Proc Am Thorac Soc* 2006, **3**(4):357-363.
35. Raghu G, Weycker D, Edelsberg J, Bradford WZ, Oster G: **Incidence and prevalence of idiopathic pulmonary fibrosis**. *Am J Respir Crit Care Med* 2006, **174**(7):810-816.
36. Pal S, Tyler JK: **Epigenetics and aging**. *Sci Adv* 2016, **2**(7):e1600584.
37. Helling BA, Gerber AN, Kadiyala V, Sasse SK, Pedersen BS, Sparks L, Nakano Y, Okamoto T, Evans CM, Yang IV *et al*: **Regulation of MUC5B Expression in Idiopathic Pulmonary Fibrosis**. *Am J Respir Cell Mol Biol* 2017, **57**(1):91-99.
38. Wang Y, Yella J, Chen J, McCormack FX, Madala SK, Jegga AG: **Unsupervised gene expression analyses identify IPF-severity correlated signatures, associated genes and biomarkers**. *BMC Pulm Med* 2017, **17**(1):133.
39. Criscione SW, Theodosakis N, Micevic G, Cornish TC, Burns KH, Neretti N, Rodic N: **Genome-wide characterization of human L1 antisense promoter-driven transcripts**. *BMC Genomics* 2016, **17**:463.
40. Lunardi F, Villano G, Perissinotto E, Agostini C, Rea F, Gnoato M, Bradaschia A, Valente M, Pontisso P, Calabrese F: **Overexpression of SERPIN B3 promotes epithelial proliferation and lung fibrosis in mice**. *Lab Invest* 2011, **91**(6):945-954.
41. Wang Z, Zhu J, Chen F, Ma L: **Weighted Gene Coexpression Network Analysis Identifies Key Genes and Pathways Associated with Idiopathic Pulmonary Fibrosis**. *Med Sci Monit* 2019, **25**:4285-4304.
42. Garcia-Perez JL, Widmann TJ, Adams IR: **The impact of transposable elements on mammalian development**. *Development* 2016, **143**(22):4101-4114.
43. Chen LL, Yang L: **ALUternative Regulation for Gene Expression**. *Trends Cell Biol* 2017, **27**(7):480-490.
44. Deininger P, Morales ME, White TB, Baddoo M, Hedges DJ, Servant G, Srivastav S, Smither ME, Concha M, DeHaro DL *et al*: **A comprehensive approach to expression of L1 loci**. *Nucleic Acids Res* 2017, **45**(5):e31.
45. Baumgartner KB, Samet JM, Stidley CA, Colby TV, Waldron JA: **Cigarette smoking: a risk factor for idiopathic pulmonary fibrosis**. *Am J Respir Crit Care Med* 1997, **155**(1):242-248.
46. Hackett NR, Butler MW, Shaykhiev R, Salit J, Omberg L, Rodriguez-Flores JL, Mezey JG, Strulovici-Barel Y, Wang G, Didon L *et al*: **RNA-Seq quantification of the human small airway epithelium transcriptome**. *BMC Genomics* 2012, **13**:82.
47. Hogg JC, Chu F, Utokaparch S, Woods R, Elliott WM, Buzatu L, Cherniack RM, Rogers RM, Sciurba FC, Coxson HO *et al*: **The nature of small-airway obstruction in chronic obstructive pulmonary disease**. *N Engl J Med* 2004, **350**(26):2645-2653.
48. Bollati V, Baccarelli A, Hou L, Bonzini M, Fustinoni S, Cavallo D, Byun HM, Jiang J, Marinelli B, Pesatori AC *et al*: **Changes in DNA methylation patterns in subjects exposed to low-dose benzene**. *Cancer Res* 2007, **67**(3):876-880.
49. Yang WR, Ardeljan D, Pacyna CN, Payer LM, Burns KH: **SQuIRE reveals locus-specific regulation of interspersed repeat expression**. *Nucleic Acids Res* 2019, **47**(5):e27.
50. Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2**. *Genome Biol* 2014, **15**(12):550.
51. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ: **The UCSC Table Browser data retrieval tool**. *Nucleic Acids Res* 2004, **32**(Database issue):D493-496.

52. Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ: **Software for computing and annotating genomic ranges**. *PLoS Comput Biol* 2013, **9**(8):e1003118.
53. Gel B, Diez-Villanueva A, Serra E, Buschbeck M, Peinado MA, Malinverni R: **regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests**. *Bioinformatics* 2016, **32**(2):289-291.
54. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features**. *Bioinformatics* 2010, **26**(6):841-842.
55. Grömping U: **Relative Importance for Linear Regression in R: The Package relaimpo.** . *Journal of Statistical Software* 2006, **17**:1–27.
56. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P *et al*: **STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets**. *Nucleic Acids Res* 2019, **47**(D1):D607-D613.
57. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, 3rd, Hao Y, Stoeckius M, Smibert P, Satija R: **Comprehensive Integration of Single-Cell Data**. *Cell* 2019, **177**(7):1888-1902 e1821.
58. Schoggins JW, Rice CM: **Interferon-stimulated genes and their antiviral effector functions**. *Curr Opin Virol* 2011, **1**(6):519-525.

Tables and Figures

Table 1. Summary of all the identified genes and TE expression changes.
Significant changes (p.adjusted < 0.005) between the donors and IPF patients in the WL, AT2 and AM are shown.

	WL	**AT2**	**AM**
TE Family	UP/DOWN	UP/DOWN	UP/DOWN
SINE	0/0	1/1	1/0
LINE	0/4	17/2	0/1
LTR	11/0	40/17	6/5
Other	3/0	14/2	4/0
Genes	654/516	2477/1654	566/467

The TE changes are shown at the family level whereas each number represents number of TEs subfamily changes. WL: whole lung; AT2: alveolar type II cell; AM: alveolar macrophage.

Table 2. Genomic distribution and gene ontology (GO) enrichment of genes significantly associated with TEs in the AT2 and AM cells.

A. GO enrichment of genes significantly related to TEs expression. A gene ontology enrichment analysis was performed on those genes with significant cis TE associations using the STRING database [56]. The total number of observed and background genes in each GO term category as well as all false discovery rate values (FDR) < 0.05 are shown for the AT2 and AM cells.

AT2				
GO TERM	**Description**	**observed gene count**	**background gene count**	**FDR**
GO:0035082	axoneme assembly	6	59	0.0022
GO:0003341	cilium movement	5	61	0.009
GO:0060271	cilium assembly	10	326	0.009
GO:0070286	axonemal dynein complex assembly	4	32	0.009
GO:0120031	plasma membrane bounded cell projection assembly	11	413	0.009
GO:0036159	inner dynein arm assembly	3	16	0.0263
GO:0070925	organelle assembly	12	666	0.044
AM				
GO:0006955	immune response	16	1560	0.0163
GO:0045124	regulation of bone resorption	4	38	0.0163
GO:0002376	immune system process	19	2370	0.0282

B. Summary of genomic distribution of TE-genes cis loci pairs in AM and AT2 cells identified by linear model test. Number of significantly observed changes are shown per each genomics region.

AT2				
TE family	**Intron**	**Intergenic**	**3'UTR3**	**5'UTR**
SINE	30	5	38	1
LINE	26	3	26	0
LTR	14	4	11	1
Other	5	0	7	1
AM				
SINE	7	0	46	3
LINE	4	0	22	0
LTR	1	0	5	0
Other	1	0	14	1

Figure 1. Changes in the expression of TE families in individual cell types and whole-lung tissue. A. Bulk RNA-seq of whole-lung tissue (WL). B. Bulk RNA-seq of flow cytometry-sorted alveolar type II cells (AT2). C. Bulk RNA-seq of flow cytometry-sorted alveolar macrophages (AM). The TE expression was determined by read counts using the SQUIRE suite of tools, while the differential expression analysis was performed using the *DESeq2* package in R. The x-axis represents the $\log_2$ ratio of the TE subfamily expression between the IPF patients and the donors. The y-axis represents the adjusted p-value based on $-\log_{10}$. The yellow dots represent the TE subfamilies with a fold expression change (FC) > 1 and p.adjust < 0.05 (as represented by two vertical dotted lines). The blue dots represent the TEs with an adjusted p-value (p.adjust) < 0.05, while the green dots represent the TEs with an FC > 1. The grey dots represent changes that do not have a significant p.adjust value or an FC higher than 1.

Figure 2. Heatmap of the correlation values between the TEs and the genes involved in TE regulation. The L1 TE subfamilies are shown on the x-axis, while the genes are shown on the y-axis. The horizontal green bar symbolizes the donor samples, while red bar symbolizes the IPF patient samples. The heatmap colors represent the correlation values between the genes and TEs from -1 (blue, negative correlation) to 1 (red, positive correlation).

Figure 3. Differentially regulated TE loci in IPF patients vs. donors in cell types and whole-lung tissue. A. Bulk RNA-seq of whole-lung tissue. B. Bulk RNA-seq of flow cytometry-sorted alveolar type II cells (AT2). C. Bulk RNA-seq of flow cytometry-sorted alveolar macrophages (AM). The TE expression was determined by read counts using the SQUIRE suite of tools, while the differential expression analysis was performed using the *DESeq2* package in R. The x-axis represents the $\log_2$ ratio of the TE locus expression between the IPF patients and donors. The y-axis represents the adjusted p-value based on $-\log_{10}$. The yellow dots represent the TE loci with a fold expression change (FC) > 1 and p.adjust < 0.05 (as represented by two vertical dotted lines). The blue dots represent the TEs with an adjusted p-value (p.adjust < 0.05), while the green dots represent the TEs with an FC > 1. The grey dots represent changes that do not have a significant p.adjust value change or an FC higher than 1.

Figure 4. Summary of the IPF-associated TE loci changes and their association with the adjacent genes. A. A Venn diagram of the overlap between the upregulated TE loci in the IPF patients when compared with the donors in whole-lung tissue (WT), alveolar type II cells (AT2), and alveolar macrophages (AM) (upper number), with the overlap between the upregulated (first number in the brackets) and downregulated genes (second number in the brackets). B. Histograms summarizing the TE expression changes by intergenic and intragenic regions in the WL, AT2 and AM. The intragenic regions are divided into the 3'UTR, 5'UTR, intron, and exon regions. The different colors represent the proportion of each TE family in each region. The x-axis represents the individual genomic regions. The y-axis represents the percentage of TE loci per TE family that are normalized by the total number of each TE family present in each genomic region.

Figure 5. Examples of differentially expressed TE loci in the AM and AT2. Tracks from the individual IPF patient samples (n = 8) and donor samples (n = 14) were overlaid on a single track. The read count histograms per base pair are shown on the y-axis and represent the collapsed counts per each phenotype group. The chromosome track represents location on the chromosome, while the axis track represents the positions on that chromosome. The annotated genes are shown on the UCSC Gene track and the annotated TEs on the Repeat Masker track (GRCh38/hg38). The highlighted region indicates that the annotated Repeat Masker TE meets the significant differential expression criteria ($\log_2$FC > 1, p.adj < 0.05). A. Intergenic locus from the L1PA3 subfamily in the AT2 cells on chromosome 18 (chr18:27279551-27285578-L1PA3:L1:LINE). B. Independent TE expression of L1PA7 in intragenic region (chr2:20458307-20464784-L1PA7:L1:LINE).

Figure 6. Barplots summarizing the TE-gene associations in the whole-lung tissue, AT2, and AM. The association between the log_2-transformed TE expression (independent variable) and the log_2-transformed gene expression (dependent variable) was tested using a linear model (lm). The p-value correction was performed using the Benjamini and Hochberg (BH) test. Only the expressions of significant genes and significantly upregulated TEs located next to the gene (max distance 50 kb) cis loci are summarized here.

Figure 7. Examples of the gene expression correlation with the adjacent TE loci. A. Correlation plots representing the correlation between the MUC5B, CHL1, SERPINB3, and GISTA2 genes and the adjacent TE loci in the AT2 cells. B. Correlation plots representing the correlation of the CCL2, MMP7, IL1RN, SPP1, and genes and the TEs in the AM cells. The x-axis represents the normalized expression values for the TEs, while the y-axis represents the normalized expression values for the genes. The colors represent the phenotypes (Green: donor; Red: IPF patient). The BH p.adjusted values for each relationship are shown; $p.adjust^1$ represents relationship between gene expression and adjacent TE-locus expression, and $p.adjust^2$ represent difference in the regression slopes between the two groups (IPF-red, and donor-green).

Figure 8. ScRNA-seq analysis summary. A. Visualization of the different cell types in a uniform manifold approximation and projection (UMAP). B. UMAP representation of the cells from donor (eight samples) and IPF (eight sample) patients. C. The number of dysregulated TE subfamilies in eight fibrosis patients vs. eight donors in different cell types. A total of 77,517 scRNA sequencing were profiled. The analyses were performed using the *Seurat* (v3) R package. The cells were clustered using a graph-based shared nearest neighbor clustering approach, and 14 cell types were identified.

Figure 9. Differential L1 score distribution in the AT2 cells and its association with cellular factors. A. The L1 score differences in the AT2 cells from eight fibrotic (green) vs. eight normal (red) lungs (p = 1.18-e09, n = 8 per phenotypic group). The L1 scores were calculated per each cell using all the expressed L1 elements, and the significance of the L1 score distribution between the phenotypes was determined using the Wilcoxon rank-sum test. The calculations were performed using the *Seurat* (v3) R package. B. Correlation plot showing the relationship between the TE L1 score and the individual TE-related cellular factors in the AT2 cells. The difference in the gene expression response to the L1 score in the AT2 cells was modeled using the *lm* function in R, the TE L1 score (independent variable), and the log_{2-} transformed gene expression (dependent variable). To determine whether the relationship between L1 score and average gene expression differs between IPF and donors we used t-test. The p-value was corrected for multiple tests using the Benjamini and Hochberg (BH) test.

Supplemental Figures

Supplemental Figure 1. Feature plot of the expressed cell markers for the cell-type identification in the UMAP plot. The cell types are classified as epithelial cells (alveolar type II cells (AT2), alveolar type I cells (AT1), ciliated cells, basal cells, and club cells), immune cells (alveolar macrophages (AM0, monocytes, B cells, plasma mast cells, dendritic cells, and T cells), and mesenchymal cells (fibroblasts and endothelial cells).

Supplemental Tables

Supplemental Table 1. Summary of the samples used in the bulk RNA-seq and scRNA-seq analyses. Each row represents the sample ID, while the columns indicate the histological types, age, sex, race, smoking history, and phenotype, as reported in the original publication [16]. All the fibrosis samples are grouped and analyzed together, as indicated in the table (analyzed phenotype. HP: hypersensitivity pneumonitis; ILD: interstitial lung disease; IPF: idiopathic pulmonary fibrosis; NA: not

available; SSc: systemic sclerosis. An additional sample description can be obtained from the original manuscript.

Supplemental Table 2. Summary of the differential analysis and correlation between the TE families and genes in AT2 cells. A. Manually curated list of genes known to be part of the L1 defense and antiviral interferon-stimulating genes [17, 58]. B. List of the significant differentially expressed TE-related genes and L1 TE families in AT2 cells. C. Correlation between 12 significant TE families and 126 significant TE-related genes. The values represent the correlation between each gene (rows) and each TE subfamily (columns) for the IPF patient and donor samples for LINE subfamilies. The plus and minus signs next to the gene names indicate strand. D. List of the gene-L1 subfamily TE pair that significantly differed between the donors and the IPFs in AT2 cells.

Supplemental Table 3. Summary of the differential analysis of the TE loci and genes between the IPF patients and the donors in different cells. A. Whole-lung (WL) tissue, B. Alveolar type II (AT2) cells. C. Alveolar macrophages (AM). The columns represent the gene expression base mean (baseMean), $\log_2$ fold change (log2FoldChange), lfcSE (log fold standard error), stat (Wald statistics; lfc/standard error), p-value, p.adj (adjusted p-value). The plus and minus signs next to the gene names indicate the gene orientation.

Supplemental Table 4. Summary of the linear model test of all the identified TE-gene loci pairs. A. AT2 linear model analysis summary. B. AM linear model analysis summary. The association between the $\log_2$-transformed TE expression (independent variable) and the $\log_2$-transformed gene expression (dependent variable) was tested using a linear model. Genes, TE locus, distance from the gene and the genomic region of the TE locus are shown. The p value^1 represents relationship between gene expression and adjacent TE-locus expression, and the p value^2 - represent difference in the regression slopes between the two groups (IPF and donor). The p-value correction was performed using the Benjamini and Hochberg (BH) test and (p.adjust^1 and p.adjust^2).

Supplemental Table 5. Summary of the differential analysis of the genes and TEs per individual cell type in the scRNA-seq. Table 5A. Cell number for each subject in each cell type. B. Differentially expressed genes in IPF vs Donor for all cell types. Avg_logFC: Average log fold change; p-value: calculated based on the Wilcoxon rank-sum test; p.adjust: adjusted p-value based on the Bonferroni correction using all the genes in the dataset; pct.1: fraction of the cells that express the gene in IPF patients; pct.2: fraction of the cells that express the gene in donors. C. Differentially expressed TE families in the IPF patient vs. donor samples for the different cell types.
D. Differences between the IPF patients and donors represented as correlation between L1 score and TE-related gene expression. A subset of 126 genes identified in RNA bulk seq were tested. A linear model was fitted for each L1 score and the gene pairs. The p-value was calculated using t-test; p.adjust is the Benjamini and Hochberg (BH)-adjusted p-value.

Figure 1

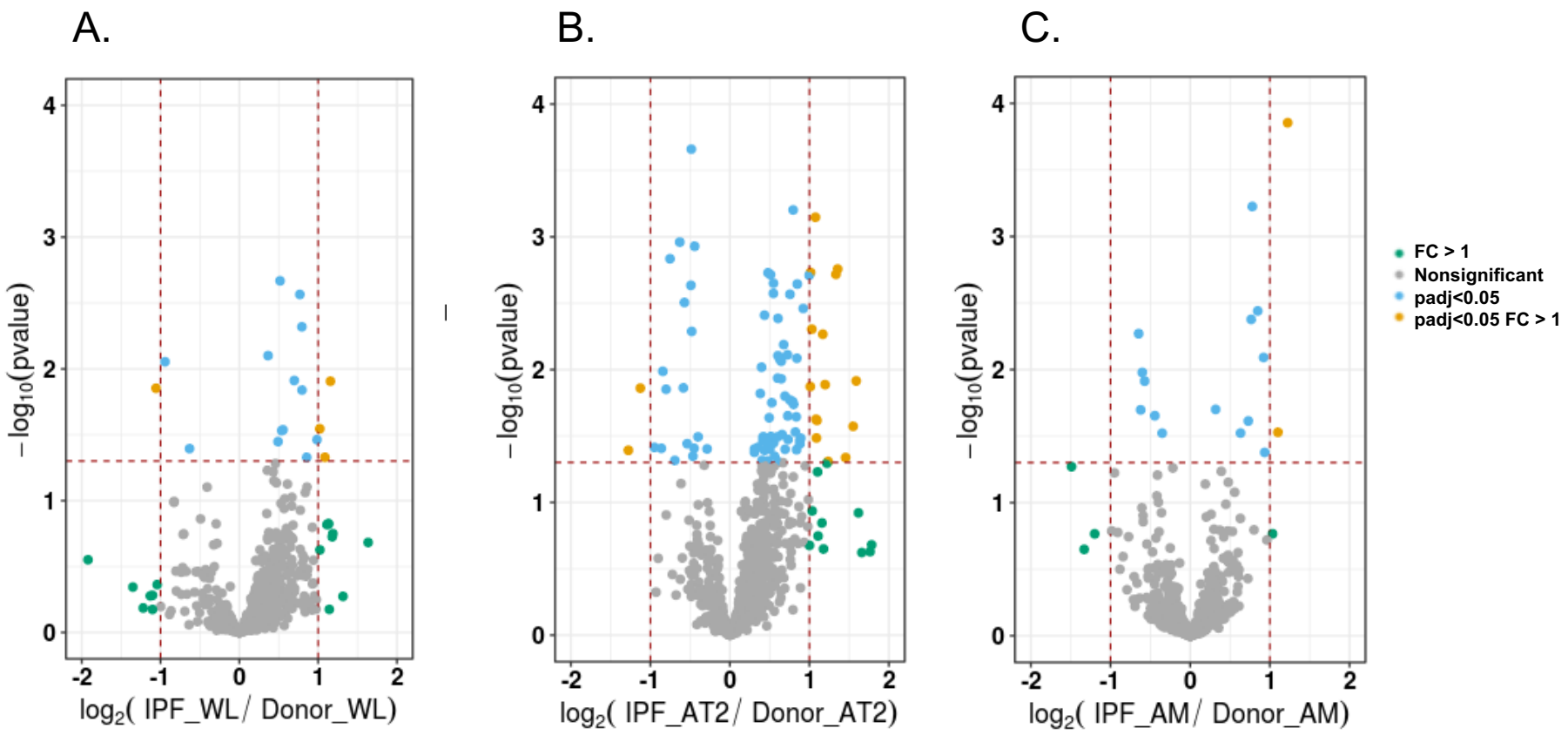

A.
B.
C.
−log10(pvalue)
log2(IPF_WL/ Donor_WL)
log2(IPF_AT2/ Donor_AT2)
log2(IPF_AM/ Donor_AM)
FC > 1
Nonsignificant
padj<0.05
padj<0.05 FC > 1

Figure 2

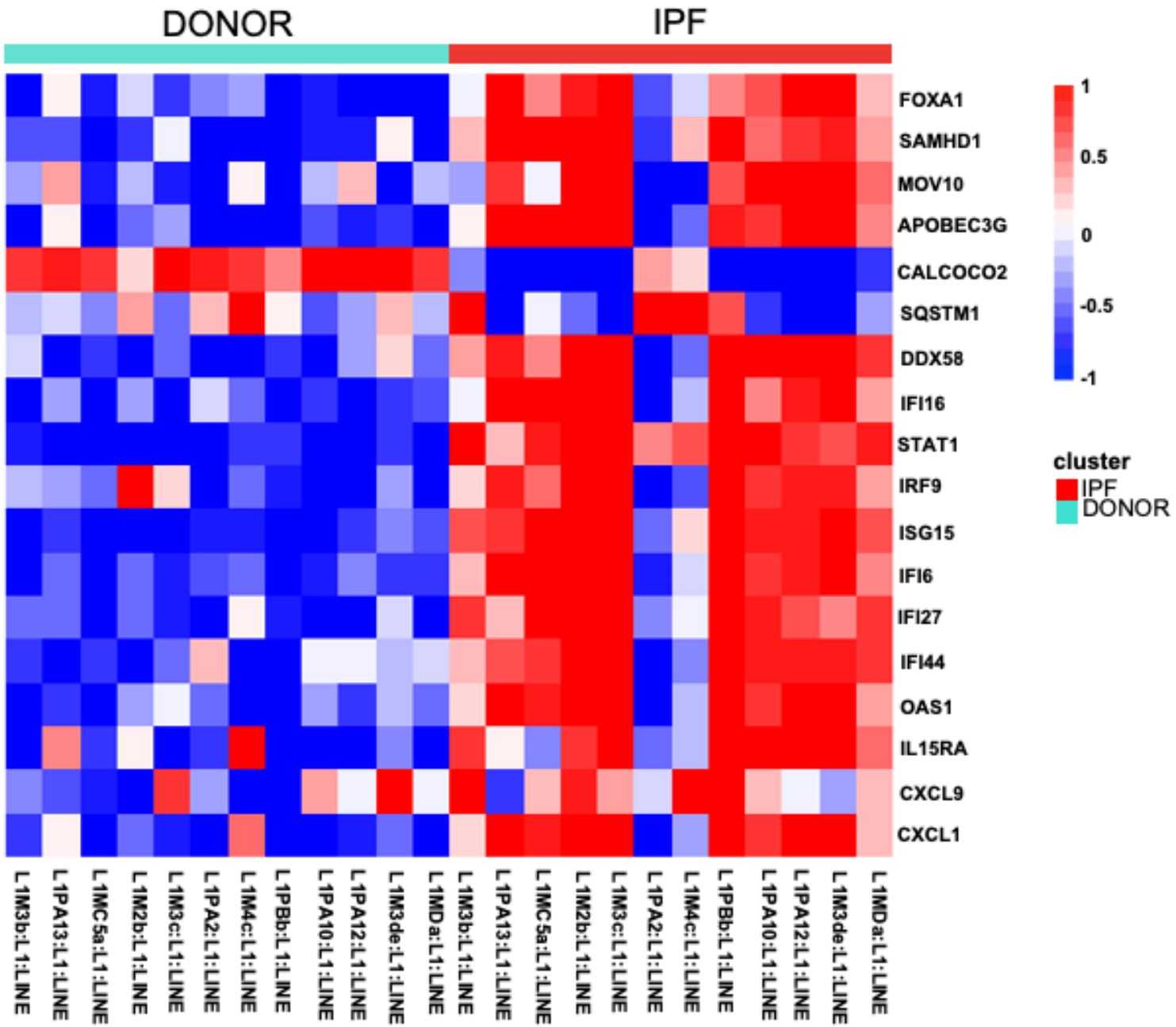

DONOR
IPF
FOXA1
SAMHD1
MOV10
APOBEC3G
CALCOCO2
SQSTM1
DDX58
IFI16
STAT1
IRF9
ISG15
IFI6
IFI27
IFI44
OAS1
IL15RA
CXCL9
CXCL1
L1M3b:L1:LINE
L1PA13:L1:LINE
L1MC5a:L1:LINE
L1M2b:L1:LINE
L1M3c:L1:LINE
L1PA2:L1:LINE
L1M4c:L1:LINE
L1PBb:L1:LINE
L1PA10:L1:LINE
L1PA12:L1:LINE
L1M3de:L1:LINE
L1MDa:L1:LINE
1
0.5
0
-0.5
-1
cluster
IPF
DONOR

Figure 3

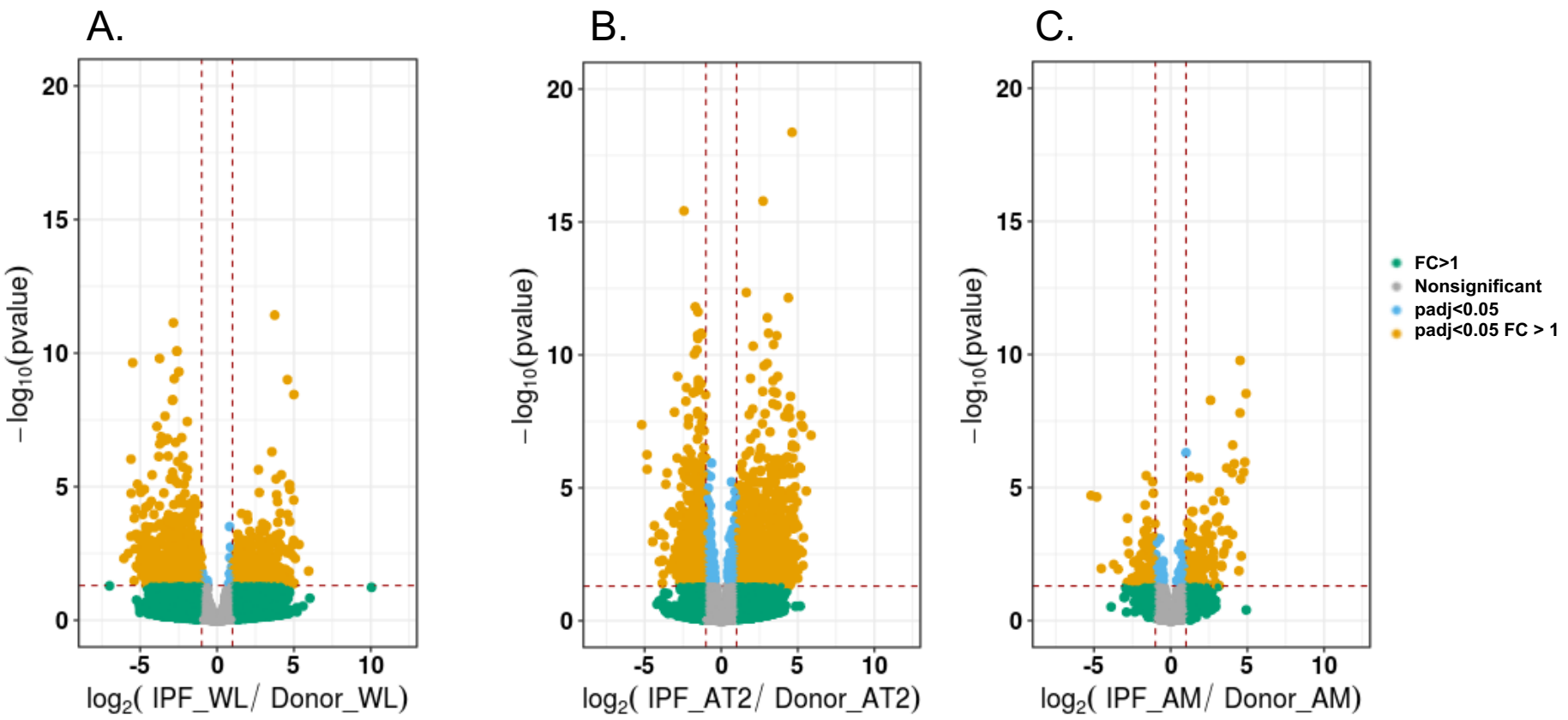
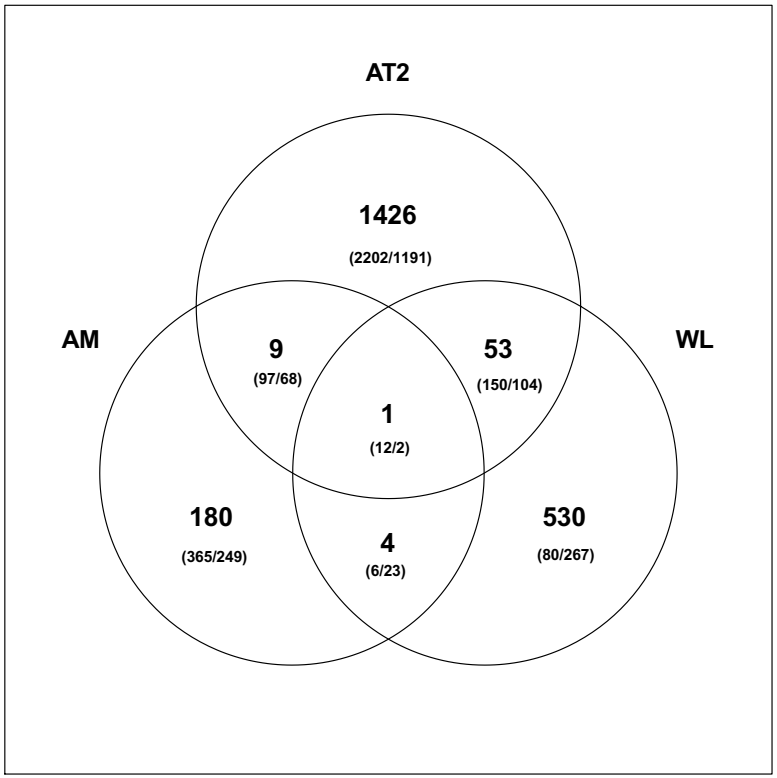

A.
B.
C.
20
15
10
5
0
-5
0
5
10
$-\log_{10}$(pvalue)
$\log_2$(IPF_WL/ Donor_WL)
$\log_2$(IPF_AT2/ Donor_AT2)
$\log_2$(IPF_AM/ Donor_AM)
FC>1
Nonsignificant
padj<0.05
padj<0.05 FC > 1

Figure 4

A.


AT2
AM
WL
1426
(2202/1191)
9
(97/68)
53
(150/104)
1
(12/2)
180
(365/249)
4
(6/23)
530
(80/267)

B.

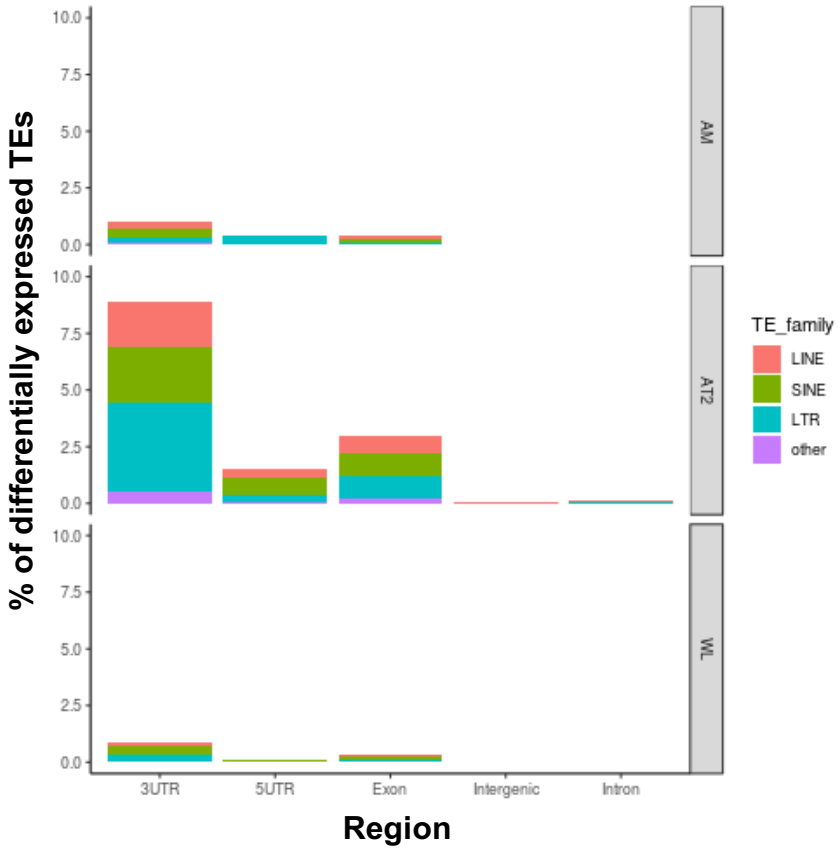
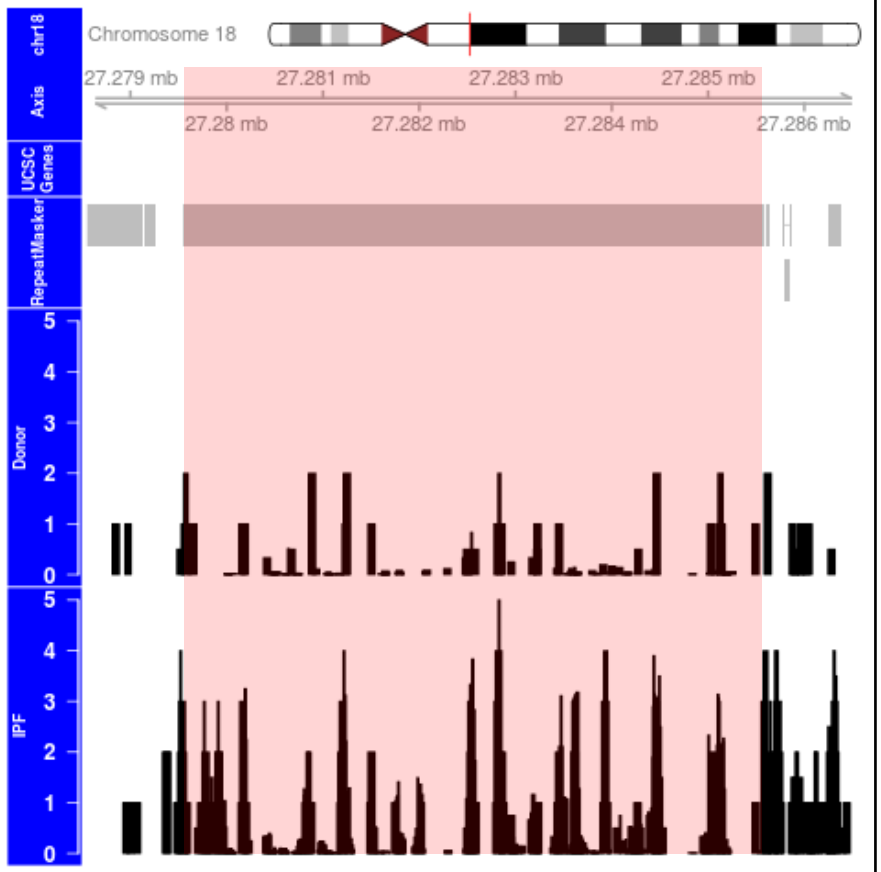

% of differentially expressed TEs
10.0
7.5
5.0
2.5
0.0
AM
AT2
WL
3UTR
5UTR
Exon
Intergenic
Intron
Region
TE_family
LINE
SINE
LTR
other

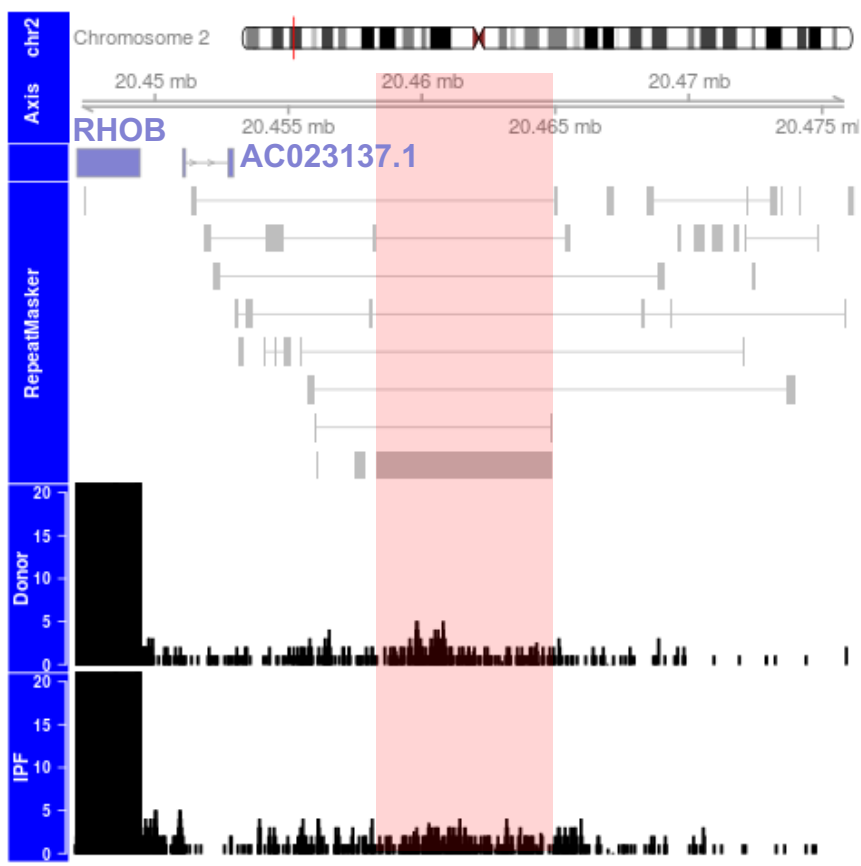
Figure 5

A.



chr18:27279551-27285578-L1PA3:L1:LINE

B.



chr2:20458307-20464784-L1PA7:L1:LINE

Figure 6

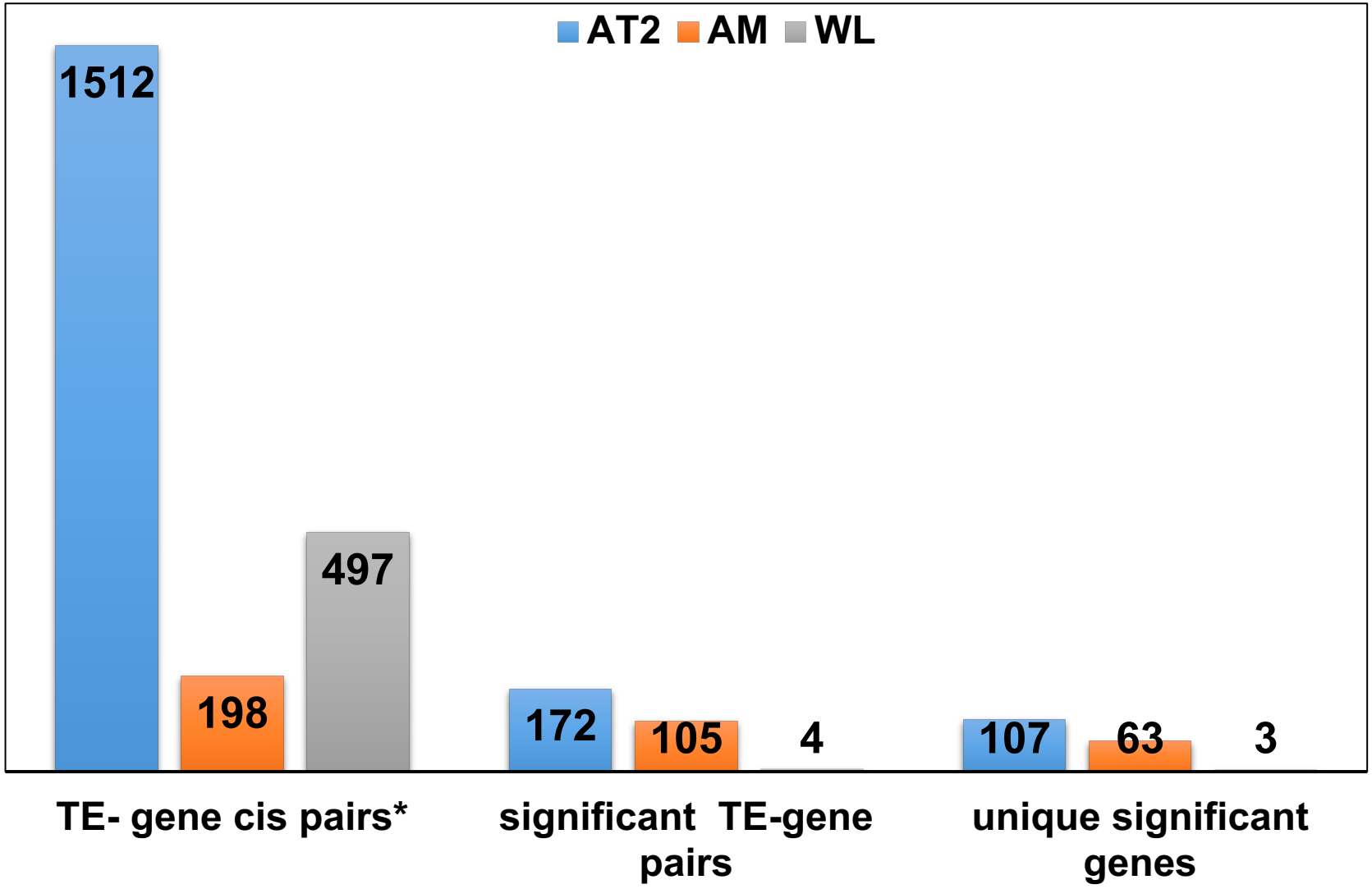

AT2
AM
WL
1512
198
497
172
105
4
107
63
3
TE- gene cis pairs*
significant TE-gene pairs
unique significant genes

Figure 7

A.

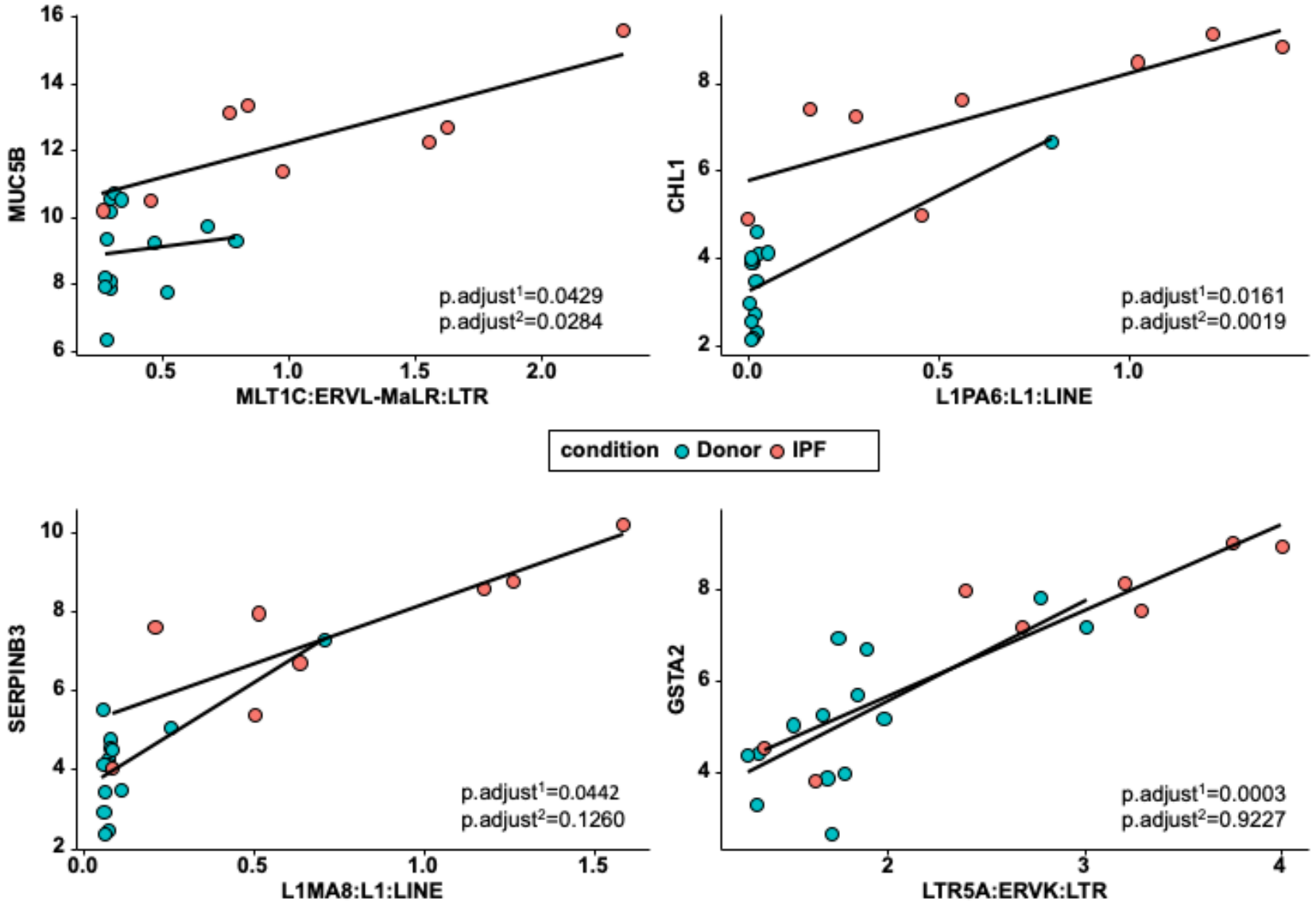

MUC5B
MLT1C:ERVL-MaLR:LTR
p.adjust[1]=0.0429
p.adjust[2]=0.0284
CHL1
L1PA6:L1:LINE
p.adjust[1]=0.0161
p.adjust[2]=0.0019
condition Donor IPF
SERPINB3
L1MA8:L1:LINE
p.adjust[1]=0.0442
p.adjust[2]=0.1260
GSTA2
LTR5A:ERVK:LTR
p.adjust[1]=0.0003
p.adjust[2]=0.9227

Figure 7

B.

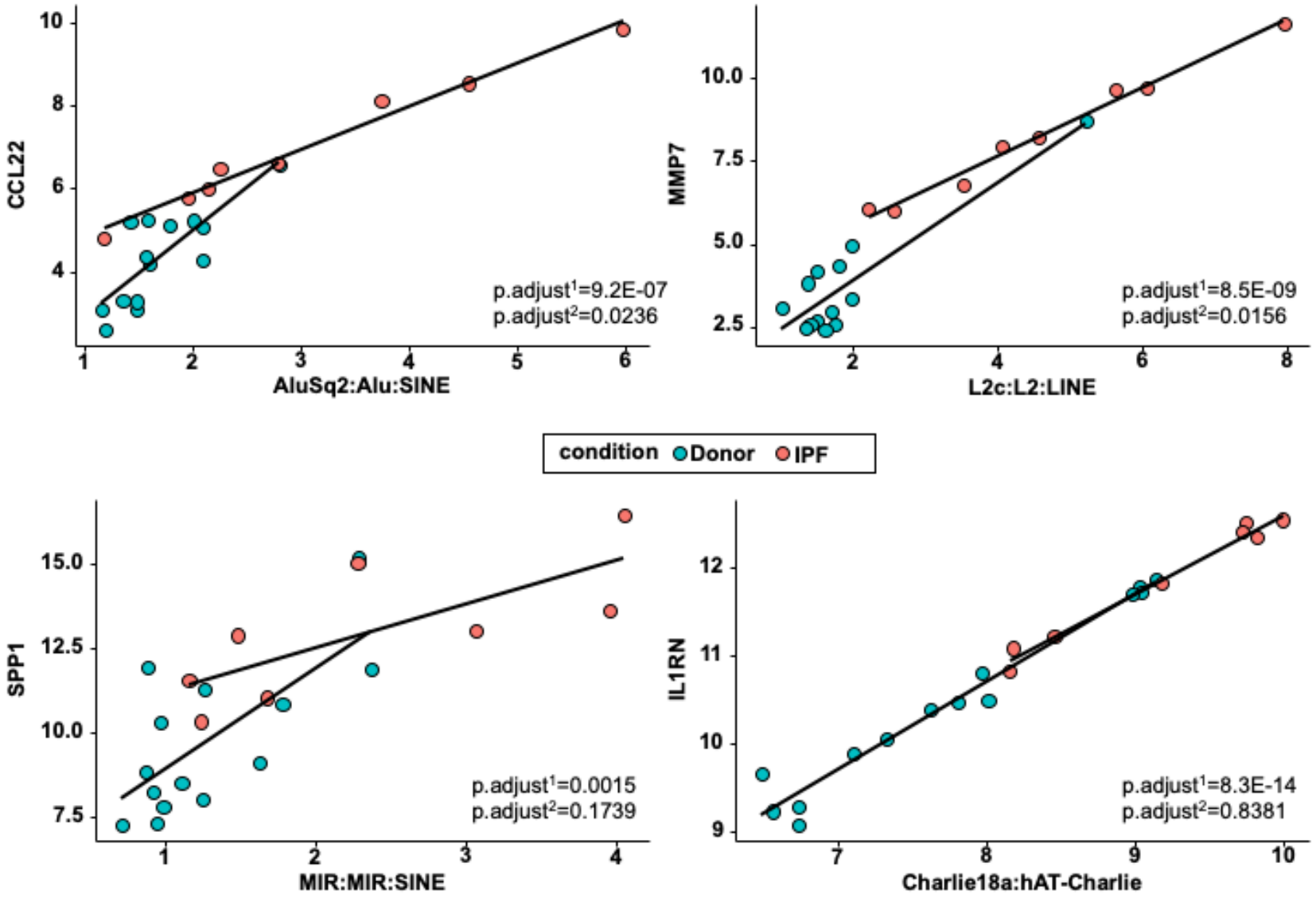

CCL22
AluSq2:Alu:SINE
p.adjust[1]=9.2E-07
p.adjust[2]=0.0236
MMP7
L2c:L2:LINE
p.adjust[1]=8.5E-09
p.adjust[2]=0.0156
condition
Donor
IPF
SPP1
MIR:MIR:SINE
p.adjust[1]=0.0015
p.adjust[2]=0.1739
IL1RN
Charlie18a:hAT-Charlie
p.adjust[1]=8.3E-14
p.adjust[2]=0.8381

Figure 8

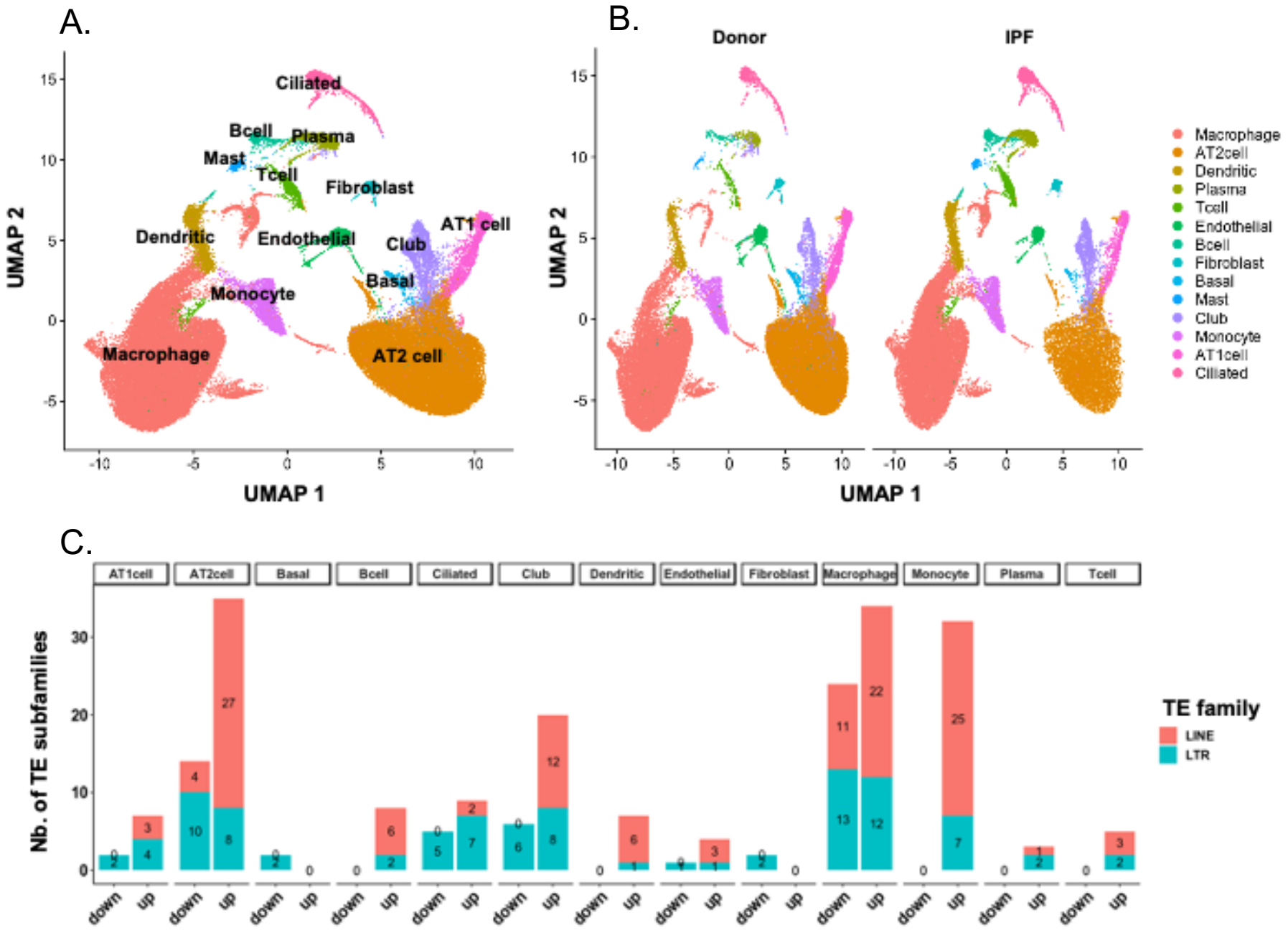
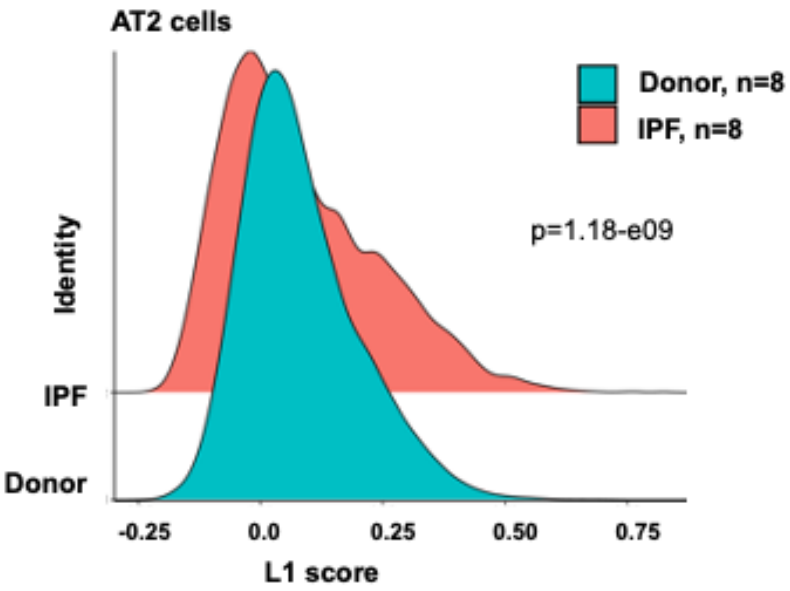
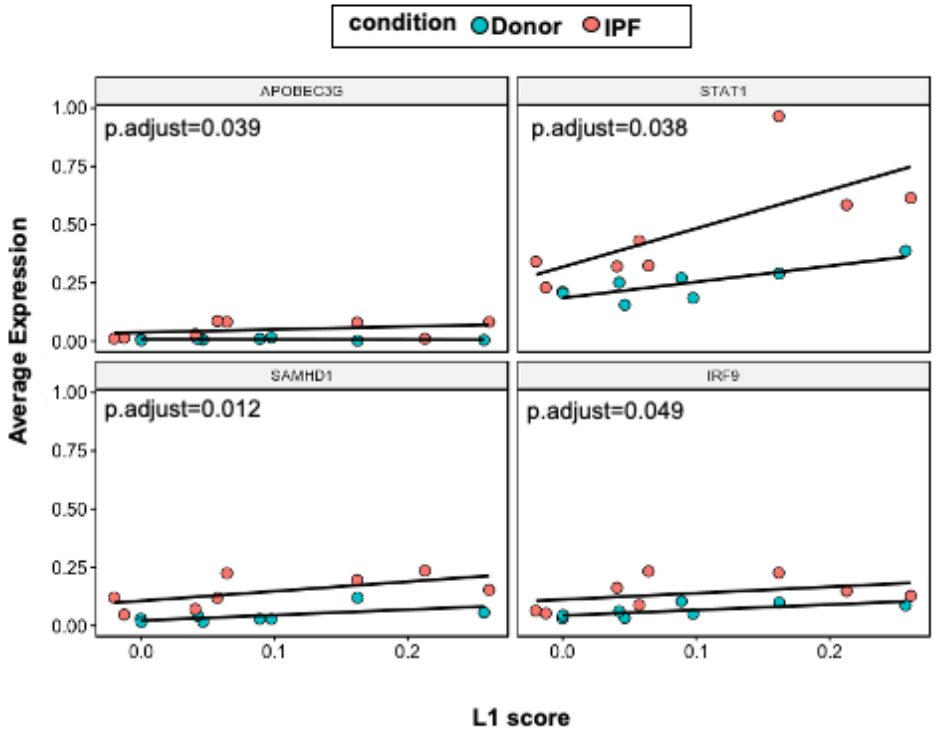

A.
B.
Donor
IPF
Ciliated
Bcell
Plasma
Mast
Tcell
Fibroblast
Dendritic
Endothelial
Club
AT1 cell
Basal
Monocyte
Macrophage
AT2 cell
UMAP 1
UMAP 2
Macrophage
AT2cell
Dendritic
Plasma
Tcell
Endothelial
Bcell
Fibroblast
Basal
Mast
Club
Monocyte
AT1cell
Ciliated
C.
AT1cell
AT2cell
Basal
Bcell
Ciliated
Club
Dendritic
Endothelial
Fibroblast
Macrophage
Monocyte
Plasma
Tcell
Nb. of TE subfamilies
down
up
TE family
LINE
LTR

Figure 9

A.


AT2 cells
Donor, n=8
IPF, n=8
p=1.18-e09
Identity
IPF
Donor
-0.25
0.0
0.25
0.50
0.75
L1 score

B.


condition
Donor
IPF
APOBEC3G
p.adjust=0.039
STAT1
p.adjust=0.038
SAMHD1
p.adjust=0.012
IRF9
p.adjust=0.049
Average Expression
1.00
0.75
0.50
0.25
0.00
0.0
0.1
0.2
L1 score