Characterizing a focused landscape of familial acute respiratory distress syndrome

Inimary Toby (itoby@udallas.edu)
university of dallas https://orcid.org/0000-0003-0820-3999
Neal Thomas
Penn State College of Medicine
Nithyananda Thorenoor
Penn State College of Medicine
Debbie Spear
Penn State College of Medicine
Susan DiAngelo
Penn State College of Medicine
Joanna Floros
Penn State College of Medicine

Research article

Keywords: Acute Respiratory Distress Syndrome, Variants, Exome sequencing, Biological pathways

Posted Date: August 21st, 2019

DOI: https://doi.org/10.21203/rs.2.13330/v1

Abstract

Background Acute respiratory distress syndrome (ARDS) affects approximately 190,600 patients per year in the United States, with mortality up to 45%. ARDS can occur as primary disease due to various factors (e.g. bacterial or viral pneumonia, gastric aspiration, lung contusion, toxic inhalation, and near drowning) or as secondary disease due to sepsis, pancreatitis, severe trauma, massive blood transfusion, and burn. We hypothesized that ARDS-affected individuals have patterns of variants in their physiological repertoire that can be tracked and utilized to complement clinical diagnosis and/or clinical monitoring. Methods The goals of this study were to: (1) characterize the landscape of variants within protein coding but we also studied UTR regions in ARDS using an Exome sequencing approach; (2), determine the variations in signaling pathways across ARDS; and (3) use computational approaches to explore the functional consequences of ARDS. Towards this we assessed an ARDS-affected individual in the context of unaffected individuals from the same family as well as unrelated ARDS cases, in order to elucidate underlying inheritance patterns of “private variants”. Private variants consist of variants shared by ARDS cases but not found in unaffected individuals. Results Whole exome sequencing yielded 3,516 variants represented by 2,354 genes. Of these, 128 variants were shared across all ARDS cases. Of these, there were 24 unique variants represented by 9 ARDS genes shared by the primary ARDS-case and unrelated individuals with ARDS. The overall genes identified and subsequent analysis, demonstrate that there are important biological pathways that distinguish ARDS cases from or from non-ARDS. These pathways include: cell-to-cell signaling interaction, cell growth and proliferation and cell morphology. The data also show a coordinated effort amongst biological processes such as liver hyperproliferation and cell death that underlie the pathogenesis of ARDS. Conclusions These in-silico discoveries demonstrate a role for private variants shared by ARDS cases to be leveraged as biomarkers for clinical diagnosis and/or monitoring of ARDS.

Background

Acute respiratory distress syndrome (ARDS) is a syndrome of hypoxic respiratory failure characterized by diffuse pulmonary infiltrates and accumulation of protein-rich pulmonary edema that cause reduction in lung compliance alveolar collapse and ventilation-perfusion mismatch [1–6]. ARDS affects approximately 190,600 patients per year in the United States, with mortality up to 45% [7]. Despite improvements in intensive care during the last fifteen years, ARDS is still the major cause of mortality and morbidity in intensive care [1,2,5–7]. In fact, ARDS therapy has seen limited progress since its initial description in 1967 and management is still largely supportive, with no established therapies targeted at the primary disease processes [8]. Accordingly, there is a need for methods of early detection [9]. There has been recent recognition of the clinical and biological heterogeneity within ARDS [10–12] that reflects our incomplete understanding of the biology of ARDS. Additional contributions to the knowledge about inheritance of ARDS and/or pathogenesis will be of great benefit in moving forward with successful clinical translation of new diagnostic, preventive, and therapeutic strategies [13–16].

ARDS occurs within one week of a known clinical insult or after worsening of respiratory symptoms. It is a consequence of various risk factors including direct (e.g., bacterial or viral pneumonia, gastric aspiration, lung contusion, toxic inhalation, and near drowning) or indirect (e.g., sepsis, pancreatitis, severe trauma, massive blood transfusion, and burn) lung injury [1–6, 17–19]. There is little knowledge on the temporal relationship between detectable inflammatory changes and the onset of increased lung density, which is the current radiographic diagnostic marker for ARDS. A better understanding of these key temporal and topographic processes may contribute to advance diagnostic and prognostic biomarkers and effective therapies [7, 20–22]. Genome sequencing studies on ARDS have concentrated on identification of plasma biomarkers that may facilitate diagnosis of ARDS which could, in theory, improve clinical care, enhance our understanding of pathophysiology, and be used to enroll more homogeneous groups of patients in clinical trials of new therapies, increasing the likelihood of detecting a treatment effect [2, 23–28].

Most recently, exome sequencing studies have been reported for ARDS as part of an outgrowth of the NHLBI's Exome Sequencing Project [29, 30]. A potential limitation of these studies is that they have rarely been conducted on families with sample collection across several generations. An advantage of exome sequencing is that it allows for the analysis of "private" gene variants—variants that may have arisen de novo in one individual or family and thus would not be detected in another [30]. We hypothesized that ARDS-affected individuals have patterns of variants expressed in their physiological repertoire that can be tracked and utilized to complement approaches to clinical diagnosis and/or clinical monitoring. To address this hypothesis, we utilized an exome sequencing approach which focuses on just the protein coding but also UTRs sequences in a given sample. Our data indicates that, there are unique variants and signaling pathways in ARDS cases which differ from those observed in unaffected individuals; and that the variant expression patterns observed in the familial cohort are markedly different from that of unrelated ARDS cases.

Results

Variants identified from a family with one ARDS case and unrelated ARDS cases.

Figure 1 shows the workflow for data collection, data filtering, and data analysis utilized in the study. The family pedigree is shown in Figure 2. The individual that was the primary focus of the study is denoted as GP7. GP1, GP2, and GP6 were the youngest family members. Variants from each sample were identified and their frequency was calculated The primary case, GP7, comprised of 3,516 highly enriched variants that were represented by 2,354 genes. A summary of the annotated functions from the full list of variants yielded the following characteristics: 343 variants were exonic non-synonymous, 8 exonic frameshift, 16 exonic non-frameshift, 356 exonic synonymous, 16 exonic unknown, 66 ncRNA exonic, 2 exonic stopgain, 118 in the 5'UTR region, 782 in the 3'UTR, 1343 were intronic/ncRNA intronic, and the remaining variants were categorized as having upstream, downstream or intergenic functions (Table 1).

Heatmaps were generated from a ranked frequency occurrence assessment from the list of variants found to be shared between GP7 and each of the GP family samples (Figure 3A) or JF unrelated ARDS samples (Figure 3B). Figure 3A illustrates the variant expression patterns and clustered relationships between GP7 and each of the GP family members. Whereas, Figure 3B illustrates the variant expression patterns between GP7 and each of the ten unrelated ARDS cases (JF9–18). The red color in each of the heatmaps denotes a presence of the same variant as found in GP7, and the green color denotes absence of the variant. Compared to the GP7 the next highest frequency of variants found amongst the family cohort was in the GP1 sample, which represented one of the younger family members (Figure 3A). This was accompanied by hierarchical clustering, which identified GP1 as the closest in variant expression pattern to GP7 (Figure 3A).

Additional hierarchical clustering assessments showed that there were 2 clades within the family cohort. The first clade was comprised of GP7, GP1, and GP4. The second clade was comprised of GP3, GP6, GP8, GP5 and GP2. The clustering profiles indicate that the most similar sample to GP7 from the second clade was GP2 (Figure 3A). Furthermore, compared to GP7, JF14 was the highest scoring sample in terms of frequency of variants present as compared to any of the other unrelated controls (Table 2). Quantification of the variant occurrence frequency showed that JF14 was ~85.5% similar to GP7 (Table 2). Clustering profiles also demonstrated that JF14 was the most similar to GP7 in variant expression pattern, thereby it was clustered closest to GP7 (Figure 3B). For the unrelated ARDS samples, JF10 had the lowest detected frequency of variants and was clustered to the farthest left of the heatmap with respect to GP7 (Table 2 and Figure 3B). The variant expression patterns show that JF10, JF16, and JF18 were the least similar to GP7, as depicted by the green colored regions denoting absence of the variant. As a result, these were all clustered to the farthest left (Figure 3B). Additional comparisons of the GP7 case and the unrelated ARDS showed that there were 128 variants shared between GP7 and all unrelated ARDS, which were represented by 104 genes (Table 3). From this 104 genes, we found there were 27 variants represented by 10 genes shared amongst GP7 and ARDS cases. Of these, there were 24 unique variants represented by 9 genes (Table 4) not found in any of the other groups shown in Table 3. The remaining 3 variants were also found in group B and were represented by the MYH14 gene.

Prediction of ARDS-related biological pathways and functions from ARDS case enriched variants

We extracted all pathway predictions for gene lists from groups (A), (B), (C) and (D) (Table 3). These were then compiled to review the statistical outcomes and determine unique pathways (Additional file 1). Variants were assessed for presence or absence in group (D) vs (A), (C) vs (A), and (B) vs (A), (Additional file 2). The outcomes from each assessment was a list comprised of variants identified along with their function and exonic information where appropriate (Additional file 2). Next, using the filtered genes from the GP7 case, we observed a ~2-fold and ~4.8-fold increase in the pool of ARDS genes for group (B) vs (A) and (C) vs (A), respectively (Table 3, Figure 4). Further assessment for each group individually based upon the presence or absence of statistically significant pathways and functions revealed one novel function, liver hyperplasia/hyperproliferation. The total number of genes found to be implicated in this function from (B), (C), and (D) was 387, 139, and 63 respectively (Figure 5). To further place this finding in

context of the overall number of genes present from each gene list, ~61% of the genes from list (D), as compared to 44% from (B) and 39% from (C) were implicated in this function (Figure 5, Additional file 1). This function was not detected in group (A) which was comprised of genes found in the ARDS literature (Additional file 1).

Pairwise assessment of the biological pathways for genes derived from groups (B), (C), and (D) as compared to (A), showed that there were 6 significant pathways for (B), namely: Cellular Effects of Sildenafil (Viagra), CCR3 Signaling in Eosinophils, Inhibition of Matrix Metalloproteases, Actin Cytoskeleton Signaling, Choline Degradation I, and PRPP Biosynthesis I and none of these were previously identified as significant pathways in (A) (Additional file 3). There were 12 significant pathways for (C) and 5 of these (Hepatic Fibrosis / Hepatic Stellate Cell Activation, Acute Phase Response Signaling, Death Receptor Signaling, Osteoarthritis Pathway, Airway Pathology in Chronic Obstructive Pulmonary Disease) were also predicted as statistically significant in (A) (Additional files 1 and 3). The other 7 significant pathways (Choline Degradation I, RhoA Signaling, ATM Signaling, Inhibition of Matrix Metalloproteases, Role of PKR in Interferon Induction and Antiviral Response, Oxidative Ethanol Degradation III, and Glutamine Biosynthesis I), were not previously identified as significant pathways in (A). There were 7 statistically significant pathways for (D) and only 1 of these (Tight junction signaling) was also predicted as statistically significant in (A) (Additional files 1 and 2), whereas the other 6 (Acyl-CoA Hydrolysis, Ceramide Biosynthesis, Actin Cytoskeleton Signaling, PTEN Signaling, Formaldehyde Oxidation II (Glutathione-dependent), and Epithelial Adherens Junction Signaling) were not. There were no pathways shared between (B), (C) and (D) as compared to A. However, there were 3 pathways shared by (B) and (C) (Inhibition of Matrix Metalloproteases, Actin Cytoskeleton Signaling, and Choline Degradation I).

Three-way pathway interaction assessments resulted in the discovery of 1 biological process that was shared by group (A), (B), and (C) but was not present in group (D). This biological process contained genes with functions in cell death and survival. We found there were 95, 123 and 237 genes implicated in this function from group (A), (B), and (C), respectively (Figure 6). Of the 95 genes from (A), 82 were shared by (A), (B), and (C) (Additional file 3). The number of additional genes contributed from (B) and (C) was 41 and 155, respectively, and this correlated positively with the increase in the gene search pool (Figure 6, Additional file 4). Additionally, we found 3 variants (Chr8 pos. 22020294 (intronic), Chr10 pos. 81316603 (3' UTR), and Chr10 pos. 81353921 (Intergenic)) from GP7 (Table 5) represented by the surfactant genes: SFTPC, SFTPA1, and SFTPA2, respectively. A search of these variants in the family and unrelated ARDS showed that there was a higher proportion of the unrelated ARDS cohort that had all 3 variants as compared to the family control cohort. Only 1 family member (GP4) had 2 out of the 3 variants (Chr8 pos. 22020294 (intronic) and Chr10 pos. 81353921 (intergenic)), whereas at least 5 of the unrelated ARDS individuals had all 3 variants.

Emergence of clusters of orthologous genes across ARDS cases

A series of data mining experiments were performed using the DAVID database resource. We conducted 3 separate experiments with genes from groups (B), (C), and (D). Genes from group (A) were excluded as these were proprietary and we did not have access to the raw file. We clustered genes from list (B) and (C) using the search parameters applied for high stringency similarity threshold (85%). We successfully matched 308 and 956 genes from gene lists (B) and (C), respectively, based upon the availability of experimental and curated information in the DAVID database. The genes from (B) were categorized into 44 clusters (Additional file 5). The genes from (C) were categorized into 92 clusters (Additional file 6). In addition, a significant majority of group (B) (i.e. 67%) and group (C) (i.e. 85%) was comprised of genes that had previously been experimentally verified and identified as having polymorphisms. For group (D), 104 genes were utilized for our initial database search and 101 genes were successfully matched to having DAVID ids.

Using the same stringency parameter as previously, the genes were categorized into 8 clusters (Additional file 7). The largest cluster for group (D) was comprised of 28% of the genes (29 genes out of 101). This cluster was annotated as having functions in transmembrane composition. Additional information from these output suggested that 71% (72) of the genes from group (D) were annotated as variants based upon previous experimental information contained in the DAVID database; 67% (i.e. 68 genes) had an identified role as polymorphisms; 42% (i.e. 42 genes) were splice variants; and 45 of these genes were annotated by DAVID as having a role in post-translational modifications based upon the inclusion criteria that the protein is post-translationally modified by the attachment of either a single phosphate group, or of a complex molecule, such as 5'-phospho-DNA, through a phosphate group. A follow up assessment using the Online Mendelian Inheritance in Man (OMIM) database, suggested potential roles for myosin light chain and myosin heavy chain genes in lung-related diseases. However, many of the other genes from groups B, C, and D revealed very little curated experimental information available in the OMIM database (Additional file 8).

Discussion

Previous genomic studies in ARDS have focused on characterization of subjects unrelated to each other in an effort to identify global genomic patterns and signatures that could be informative for diagnosis or prognosis [20, 22, 29–31]. Numerous studies have suggested that there is a delicate relationship between the gene and local environment in ARDS pathogenesis. It has also been noted that there is a physiological balancing act that plays a role in this dynamic process [3, 4, 7, 14, 20, 32]. The relationship is such that the following two events must occur to give rise to the disease: First, an individual must have a genetic propensity and second their microenvironment must undergo a struggle to execute an appropriate response to local insult such as trauma, sepsis, or infectious trigger [1–6, 17–19]. It is the combinatorial effect from these two events that preferentially activates the disease. We hypothesized that ARDS-affected individuals have patterns of variants in their physiological repertoire that can be tracked, and then these would complement clinical diagnosis and/or clinical monitoring. The goals here were (1) characterize the landscape of variants within protein coding but we also studied UTR regions in ARDS using an Exome sequencing approach; (2), determine the variations in signaling pathways across

ARDS; and (3) use computational approaches to explore the functional consequences of ARDS. The findings from this study showed that: (1) there are unique variants not found in unaffected individuals but shared by ARDS cases; (2) Coordinated signaling pathways shared by ARDS cases are different from those found in unaffected individuals; (3) Clustering analysis demonstrated that the GP7 case exhibited less similarity in variant expression when compared with related family members in contrast to higher similarity observed with ARDS unrelated individuals; (4) Validation of variants from ARDS cases represent opportunities for contribution to gaps in coverage within the ARDS literature (i.e. literature curated genes).

The strategy undertaken in this study was to focus on a structured landscape. The structured landscape in this case was defined by the shared variant inheritance pattern represented across ARDS cases, and it consisted of two components: private variants and public variants. For ARDS, we believe the interaction between variant inheritance pattern and diagnosis defines the ARDS landscape in part. The primary motivation was to elucidate underlying inheritance patterns of “private variants” that could be hidden within a larger cohort. Private variants consist of variants shared amongst ARDS cases but not found in the cohort of related family members. This is in contrast to public variants, which would consist of variants shared by ARDS cases and not present in other individuals. The question posed was, does a focused landscape of ARDS display a distinct signature from person to person or is there a shared signature? An important consideration for this study, was that there was a documented presence of ARDS based on family history. This past medical history could be tracked across a generation as demonstrated by the family pedigree from Figure 2.

The variants identified support our hypothesis that there are important biological pathways that could help distinguish ARDS cases from one another. In order to place these variants in the context of the available curated literature, it was important to parse out subsets of gene lists from the larger data. We observed that there was an increase in significant pathways predicted for the GP7 case when the search window of genes was increased by the addition of groups (B) and (C) genes to the ARDS-literature curated genes. The 3,516 variants from the study contributed a much more robust pool of candidate genes to utilize in performing pathway predictions. The increase in sample size (i.e. # of genes included in search criteria) contributed to an increase in detection power. This was due to the signal-to-detection ratio becoming much higher, thereby leading to observed increases in significant pathways predicted. The groups (B) and (C) genes served as representative subsets of the GP7 enriched variants. By using these subsets of variants, we captured a novel pathway prediction: liver hyperplasia and hyperproliferation, which we hadn’t previously observed when searching solely within group A. Liver hyperplasia and hyper-proliferation is widely implicated in respiratory complications [33–36] and has also been documented in cases of secondary ARDS due to conditions such as acute pancreatitis [37, 38]. Follow up studies would be important to better understand if the genes within this pathway persist broadly across other ARDS cases. In an effort to characterize variant patterns across all ARDS cases from this study, we parsed out the group (D) genes. The group (D) genes represented an important subset because it represented a much more condensed list of variants shared across ARDS cases (i.e. GP7 and all unrelated ARDS JF cases). Using group (D), we intentionally isolated ARDS-specific pathways that may have been missed while

searching the general pool from the other gene lists. Group (D) assessments also enabled us to place in context contributions of the larger subsets of variants from groups (B) and (C). The 9 unique genes represented by 24 variants from (D) offer potential candidates for validation. Additional identification of the 3 surfactant genes in GP7 and in the unrelated ARDS cohort was not evident in the family cohort as there was only 1 family member that had 2 out of the 3 gene variants. Surfactant genes have been implicated in ARDS across multiple studies and our finding is in line with their documented role [39–42].

Pathway assessments demonstrated that with the suppression of background noise and subsequent screening of variants highly enriched in the GP7 case, we observed some novel pathways that previously weren't recognized as statistically significant for ARDS such as C-C chemokine receptor type 3 (CCR3) signaling, phosphoribosyl pyrophosphate (PRPP) biosynthesis, inhibition of matrix metalloproteases, choline degradation, glutamine biosynthesis, actin cytoskeleton signaling and epithelial junction signaling. These molecules have documented roles in lung pathogenesis and physiological signaling [24, 27, 43–45]. This indicates that there are potentially several signaling molecules derived from GP7 that could be prevalent due to the disease process. Of the 2,354 genes that were highly enriched in the GP7 case, we were able to confirm some overlap with the preexisting ones in IPA based upon our assessments with gene list A. The observed increases in the pool of ARDS genes from gene lists B (~2-fold) and C (~4.8-fold) is remarkable and contributed to the identification of these novel pathways. Our data also confirm the presence of shared pathways with group A and GP7 enriched gene subsets, including many implicated in lung pathogenesis such as airway pathology, acute phase response signaling, and hepatic fibrosis. These findings are in line with current information on ARDS from the available curated literature genes [26, 45–50].

Interestingly, the family heatmap comparisons illustrate that the second highest frequency of variants found was in the GP1 sample, representing one of the younger family members. This finding was strengthened by hierarchical clustering analysis shown in Figure 3A, which identified GP1 as having the highest similarity in variant expression pattern to GP7. This was a particularly curious finding because GP1 was the youngest family member, and had never been diagnosed with ARDS [15, 51]. Recent findings indicate that younger persons can also get ARDS [52]. However, it is unclear what the consequences are for the variant inheritance pattern we observed in GP1 as this individual has not been diagnosed with ARDS, however further studies would be important so as to better determine what this discovery means. Additionally, a closer examination of the clade structure based upon clustering analysis shown in Figure 3A, displays a vastly different image from that of Figure 3B. To quantify the differences observed in Figure 3B, 7 out of the 10 ARDS controls had a high similarity in variant expression profiles. Whereas, 3 out of the 10 ARDS controls exhibited a much different variant expression profile as demonstrated in the heatmap assessments. Thus, the contrast between Figure 3A and 3B is such that there was a higher proportion of variants shared amongst the unrelated ARDS than within the family cohort. This indicates that basal expression patterns of variants exist, and these preferentially shared amongst ARDS cases. This presents an area of interest in order to better understand in the future whether this is predictive of clinical outcomes or disease severity for the outlier samples (i.e. samples clustered farther away from

GP7 thereby denoting less similarity). Additional speculation about the clinical implications of this current observation however, is beyond the scope of this study.

In a step towards developing a more generalized model, we applied clustering analysis for relationships between the groups of ARDS genes. The assessment of clusters of orthologous genes (COGS) categories yielded pre-existing experimental information about the genes. The DAVID resource provided a streamlined approach for identifying the roles of previously discovered variants from ARDS-related genes, which was of benefit in our classification of the variants. Functional annotations retrieved showed that much of the genes, such as the myosin light chain genes, from groups A, B, and C had previously been identified as containing polymorphisms. These data are in line with previously published information while also highlight new possibilities for exploratory work and data sharing with the broader scientific community [53]. ARDS studies done on families with previous generations are very few [27]. As the genes identified carry specific variants, it will be important to understand the relationship of these variants across multiple ARDS cases. Based upon our findings, it is important to validate the variants whose function is currently unknown. The OMIM database, which is the primary inclusion criteria for the OMIM ontology, had a lack of enough experimental information on genes from groups B, C and D as relates to ARDS,. This represents a gap in coverage within the literature for which one could contribute to experimental knowledge by validation and deposition of the variants from this study.

Conclusions

Taken together, the findings from this study promote the idea that there is a coordinated effort amongst signaling processes that underlie the pathogenesis of ARDS. This combinatorial signaling behavior has been well documented for lung pathogenesis and is implicated in ARDS. The study population represents an important demographic in helping understand the disease prevalence within a specific family in comparison to the unrelated ARDS controls. That said, the frequency of individuals sampled in the family cohort though not a large number, does show that there are some family members with a high degree of shared variant expression as compared to the primary case, GP7. We speculate that more work is needed to validate these additional variants and place them in the larger context of familial ARDS cases. The potential outcomes would contribute to efforts geared towards clinical diagnosis and/or monitoring of cases for which family history indicates the presence of a genetic inheritance pattern of ARDS.

Methods

Study information and Sample Collection

Family members and control cohort were recruited, enrolled after informed consent under a protocol approved by the Human Subjects Protection Office and Institutional Review Board from the Pennsylvania State University College of Medicine. Three sets of samples were collected: primary ARDS case, and related family members; a third set of samples from unrelated ARDS subjects used in the present study were collected as previously described [40, 41]. Together, we analyzed a total of 18 samples.

Whole Exome sequencing analysis and data collection

The workflow in Figure 1 outlines the key steps involved in sample data collection, data filtering and the downstream data analysis processes that we applied to the sequencing output in order to perform comparisons. Variants for each sample were identified based on the GATK [54] best practices pipeline. The bwa v0.7.3a software were used to align the paired end exome sequences to the hg19 reference and the Picard v1.102 Mark Duplicates tool was used to remove duplicates. Local realignment around indels was performed by running the GATK Realigner Target Creator and Indel Realigner tools, using the Mills and 1000G Gold Standard indels as the known indels. Base quality score recalibration was performed using the GATK Base Recalibrator tool with dbSNP build 138 and the Mills and 1000G Gold Standard indels as known sites, followed by GATK PrintReads. Variants were called using the GATK Haplotype Caller tools with the following parameters: ERC GVCF, LINEAR variant index type and 128000 variant index parameter followed by the GATK Joint Genotyping tool. The ANNOVAR v2015–03–22 [55] was used to functionally annotate the genetic variants, including when applicable gene membership (e.g. intron), conservation, nonsynonymous amino acid substitution, SIFT prediction, Polyphen2 prediction, etc.

Filtering and Classification of Variants

Ingenuity Pathways Analysis (IPA) was used to determine genes associated with ARDS in the literature, and to identify significantly enriched canonical pathways, networks, diseases and biological functions and upstream regulators from amongst filtered list of variants [56]. For IPA, we applied a filter to extract the most promising variants based on one case (GP7), samples from the family, and ten unrelated ARDS cases, (JF9–18) as well as control samples GP3–5 and GP8 from the same family as the case (GP7). While samples GP1, GP2, and GP6 were disease free, we excluded them from our controls since these originated from younger individuals and we were focused on late onset of ARDS during adulthood. However, these samples were later utilized in all follow-up analysis done on family variant inheritance patterns. For IPA assessments, the list of pre-processed variants was further filtered based upon the following set of criterias: (1) that GP7, which was the primary case in the family samples, had the variant; (2) at least 2 of the reduced (n = 4) family controls didn't have the variant; (3) at least 2 of the ten (JF 9–18) samples had the variant and; (4) that there were at least twice as many unrelated cases with the variant than without the variant. This process enabled us to capture the variants that were enriched in our case sample, GP7, which were represented across 16 gene/exonic functional features. From this post-processed list of variants, functional subsets were derived for further analysis. The variants were categorized according to their respective genes and were subsequently placed in groups. They were grouped from A-D, based upon the following criteria: (1) Group A- comprised of literature genes curated from ARDS-related studies currently available in IPA; (2) Group B- comprised of variants represented by exonic genes affecting stop codon, genes with frameshift mutations and non-synonymous mutations; (3) Group C- comprised of variants represented by genes (which included all genes from Group B) plus genes found within the 3' UTR, 5' UTR, and non-coding RNAs and; (4) Group D- consisted of variants that were

represented by genes shared by GP7 and all JF unrelated ARDS cases from this study. IPA analysis was performed for each group of genes.

Statistical metrics applied to pathway analysis

IPA [56] applies statistical assessments to determine pathway relationships for a given list of genes. To apply the underlying statistical metrics for a given dataset, we first uploaded a list of genes and performed a Core Analysis with the default settings in IPA. The Canonical Pathway Analysis in IPA associates the genes with the canonical pathways in Ingenuity's Knowledge Base and returns two measures of association: (1) a ratio of the number of genes from the list that maps to the pathway divided by the total number of genes that map to the same pathway and; (2) a p-value of the Fisher's exact test. To identify ARDS-specific significant gene sets that were within the top scoring targets from pathway assessments of the list of genes, we examined the top pathways from the following five IPA analysis modules: (1) canonical pathways; (2) upstream regulators; (3) diseases and bio functions; (4) Toxicity (tox) functions and; (5) networks. A Bonferroni correction (i.e. p-value < 0.05/25 = 0.002) was applied to the p-values in order to maintain the type I error at 5%. For the top gene sets obtained from the query gene list, we kept genes with consistent empirical and biological relationships. The biological relationships between genes in a canonical pathway are referred to as IPA pathways. Using the biological relationships from the IPA pathway and the top genes in the canonical pathway as references, we derived a correlation trend between 'within patient expression changes' (WPEC) and 'ordered categorical Multiple Organ Failure' (ocMOF) labels that were biologically driven for all genes in the same pathway [23]. If this trend is consistent with the one computed from the data, the gene is retained and used to compute the dominant trajectory. The IPA pathway provides a graphical representation of the biological relationships between genes in a canonical pathway, where nodes represent genes and edges represent the biological relationships. Each edge is supported by at least one reference from the literature, a textbook, or canonical information stored in the Ingenuity Pathways Knowledge Base, providing us a relationship summary between genes. For this study, a -log p value of 1.30 represents the cutoff for significance; any pathways with values less than this would not be considered as significant across the dataset.

Clustering of variants and heatmap assessments

The filtered variants were extracted and parsed for all samples from the study. These tables were then imported into an excel spreadsheet and calculations were done for the presence (assigned a 1) or absence (assigned a 0) of variants within a specific sample using VB scripting [57]. The final csv file was uploaded into R [58]. We then applied the 'Heatmap' and 'clustergram' scripts using R packages gplots, gdata, gtools, and rcolorbrewer to render a 2-d color image of the data showing the samples on the x-axis. To organize these data and identify potential relationships among presence/absence of specific variants, we utilized a hierarchical clustering with Euclidean distance metric and average linkage to generate the hierarchical tree. This type of clustering enabled us to find the similarity or dissimilarity between every pair of objects in the data set, group the objects into a binary, hierarchical cluster tree, and determine where to differentiate the hierarchical tree into clusters.

Clustering of genes and annotation analysis

DAVID (Database for Annotation, Visualization, and Integrated Discovery) is a Web-based application that provides a high-throughput and integrative gene functional annotation environment to systematically extract biological themes behind large gene lists. High-throughput gene functional analysis with DAVID helps to provide important insights that allow investigators to understand the biological themes within their given genomic study [59–61]. We took gene lists from groups: (B), (C) and (D) (Table 2) and performed clustering analysis using the DAVID tool. We were unable to perform this type of analysis for group A, as this gene list is proprietary information and is retained by Agilent Biotechnologies. We performed all searches using the default parameters, as referenced in the software manual, based upon the “highest” classification stringency cutoff to obtain functionally related gene groups. For the annotation analysis generated from DAVID, we searched all human sequence related metadata available in the database and we then extracted the data to explore gene cluster relationships for each of our gene lists. DAVID uses a set of fuzzy classification algorithms to group genes based on their co-occurrences in annotation terms and ranks the gene groups using an internal (EASE) score [62].

List Of Abbreviations

Tox- Toxicity, DAVID- Database for Annotation, Visualization, and Integrated Discovery, OMIM-Online Mendelian Inheritance in Man, and IPA- Ingenuity pathway analysis

Declarations

Ethics approval and consent to participate

The protocol approved by the Human Subjects Protection Office and Institutional Review Board from the Pennsylvania State University College of Medicine.

Consent for publication

Not applicable.

Availability of data and material

The datasets used and/or analyzed during the current study are available from the corresponding authors on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

This work is supported by Dr. John Ardell Pursley Memorial Research Fund and Pedlow’s research gift.

Authors' contributions

IT performed the analyses and wrote the paper, NJT wrote and was responsible for the human subjects approval protocol, NT performed IPAs and helped with manuscript preparation; DS collected samples from family, SD extracted DNA, and handled the samples, JF oversaw all aspects of the project, worked with IT in aspects of data analysis and integration and contributed to the preparation of the manuscript.

Acknowledgements

The authors wish to thank Ms. Anna Salzberg for her help with some of the analysis steps.

References

1.Force ADT, Ranieri VM, Rubenfeld GD, Thompson BT, Ferguson ND, Caldwell E, et al. Acute respiratory distress syndrome: the Berlin Definition. JAMA. 2012;307(23):2526–33.

2.Ware LB, Matthay MA. The acute respiratory distress syndrome. N Engl J Med. 2000;342(18):1334–49.

3.Phua J, Badia JR, Adhikari NK, Friedrich JO, Fowler RA, Singh JM, et al. Has mortality from acute respiratory distress syndrome decreased over time?: A systematic review. Am J Respir Crit Care Med. 2009;179(3):220–7.

4.Rubenfeld GD, Caldwell E, Peabody E, Weaver J, Martin DP, Neff M, et al. Incidence and outcomes of acute lung injury. N Engl J Med. 2005;353(16):1685–93.

5.Katzenstein AL, Bloor CM, Leibow AA. Diffuse alveolar damage—the role of oxygen, shock, and related factors. A review. Am J Pathol. 1976;85(1):209–28.

6.Matute-Bello G, Frevert CW, Martin TR. Animal models of acute lung injury. Am J Physiol Lung Cell Mol Physiol. 2008;295(3):L379–99.

7.Wellman TJ, de Prost N, Tucci M, Winkler T, Baron RM, Filipczak P, et al. Lung Metabolic Activation as an Early Biomarker of Acute Respiratory Distress Syndrome and Local Gene Expression Heterogeneity. Anesthesiology. 2016;125(5):992–1004.

8.Ashbaugh DG, Bigelow DB, Petty TL, Levine BE. Acute respiratory distress in adults. Lancet. 1967;2(7511):319–23.

9.Janz DR, Ware LB. The needle in the haystack: searching for biomarkers in acute respiratory distress syndrome. Crit Care. 2013;17(5):192.

10.Dowdy DW, Eid MP, Dennison CR, Mendez-Tellez PA, Herridge MS, Guallar E, et al. Quality of life after acute respiratory distress syndrome: a meta-analysis. Intensive Care Med. 2006;32(8):1115–24.

11.Yehya N, Thomas NJ, Wong HR. Evidence of Endotypes in Pediatric Acute Hypoxemic Respiratory Failure Caused by Sepsis. Pediatr Crit Care Med. 2019;20(2):110–2.

12.Sweeney TE, Thomas NJ, Howrylak JA, Wong HR, Rogers AJ, Khatri P. Multicohort Analysis of Whole-Blood Gene Expression Data Does Not Form a Robust Diagnostic for Acute Respiratory Distress Syndrome. Crit Care Med. 2018;46(2):244–51.

13.Vincent JL, Sakr Y, Sprung CL, Ranieri VM, Reinhart K, Gerlach H, et al. Sepsis in European intensive care units: results of the SOAP study. Crit Care Med. 2006;34(2):344–53.

14.Tejera P, Meyer NJ, Chen F, Feng R, Zhao Y, O'Mahony DS, et al. Distinct and replicable genetic risk factors for acute respiratory distress syndrome of pulmonary or extrapulmonary origin. J Med Genet. 2012;49(11):671–80.

15.Chiumello D, Marino A. ARDS onset time and prognosis: is it a turtle and rabbit race? J Thorac Dis. 2017;9(4):973–5.

16.Constantin JM, Grasso S, Chanques G, Aufort S, Futier E, Sebbane M, et al. Lung morphology predicts response to recruitment maneuver in patients with acute respiratory distress syndrome. Crit Care Med. 2010;38(4):1108–17.

17.Alberti C, Brun-Buisson C, Goodman SV, Guidici D, Granton J, Moreno R, et al. Influence of systemic inflammatory response syndrome and sepsis on outcome of critically ill infected patients. Am J Respir Crit Care Med. 2003;168(1):77–84.

18.Brun-Buisson C, Minelli C, Bertolini G, Brazzi L, Pimentel J, Lewandowski K, et al. Epidemiology and outcome of acute lung injury in European intensive care units. Results from the ALIVE study. Intensive Care Med. 2004;30(1):51–61.

19.Calfee CS, Janz DR, Bernard GR, May AK, Kangelaris KN, Matthay MA, et al. Distinct molecular phenotypes of direct vs indirect ARDS in single-center and multicenter studies. Chest. 2015;147(6):1539–48.

20.Blondonnet R, Constantin JM, Sapin V, Jabaudon M. A Pathophysiologic Approach to Biomarkers in Acute Respiratory Distress Syndrome. Dis Markers. 2016;2016:3501373.

21.Puybasset L, Cluzel P, Chao N, Slutsky AS, Coriat P, Rouby JJ. A computed tomography scan assessment of regional lung volume in acute lung injury. The CT Scan ARDS Study Group. Am J Respir Crit Care Med. 1998;158(5 Pt 1):1644–55.

22.Walter JM, Wilson J, Ware LB. Biomarkers in acute respiratory distress syndrome: from pathobiology to improving patient care. Expert Rev Respir Med. 2014;8(5):573–86.

23.Desai KH, Tan CS, Leek JT, Maier RV, Tompkins RG, Storey JD, et al. Dissecting inflammatory complications in critically injured patients by within-patient gene expression changes: a longitudinal clinical genomics study. PLoS Med. 2011;8(9):e1001093.

24.Matthay MA, Zemans RL. The acute respiratory distress syndrome: pathogenesis and treatment. Annu Rev Pathol. 2011;6:147–63.

25.Meduri GU, Annane D, Chrousos GP, Marik PE, Sinclair SE. Activation and regulation of systemic inflammation in ARDS: rationale for prolonged glucocorticoid therapy. Chest. 2009;136(6):1631–43.

26.Nick JA, Caceres SM, Kret JE, Poch KR, Strand M, Faino AV, et al. Extremes of Interferon-Stimulated Gene Expression Associate with Worse Outcomes in the Acute Respiratory Distress Syndrome. PLoS One. 2016;11(9):e0162490.

27.Reilly JP, Christie JD, Meyer NJ. Fifty Years of Research in ARDS. Genomic Contributions and Opportunities. Am J Respir Crit Care Med. 2017;196(9):1113–21.

28.Ren S, Chen X, Jiang L, Zhu B, Jiang Q, Xi X. Deleted in malignant brain tumors 1 protein is a potential biomarker of acute respiratory distress syndrome induced by pneumonia. Biochem Biophys Res Commun. 2016;478(3):1344–9.

29.Shortt K, Chaudhary S, Grigoryev D, Heruth DP, Venkitachalam L, Zhang LQ, et al. Identification of novel single nucleotide polymorphisms associated with acute respiratory distress syndrome by exome-seq. PLoS One. 2014;9(11):e111953.

30.Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science. 2012;337(6090):64–9.

31.Ware LB, Koyama T, Zhao Z, Janz DR, Wickersham N, Bernard GR, et al. Biomarkers of lung epithelial injury and inflammation distinguish severe sepsis patients with acute respiratory distress syndrome. Crit Care. 2013;17(5):R253.

32.Matthay MA, Ware LB, Zimmerman GA. The acute respiratory distress syndrome. J Clin Invest. 2012;122(8):2731–40.

33.Bryant BH, Zenali MJ, Swanson PE, Upton MP, Yeh MM, Cuevas C, et al. Glutamine Synthetase Immunoreactivity in Peritumoral Hyperplasia in Liver: Case Report of a Metastatic Paraganglioma With Focal Nodular Hyperplasia-Like Changes and Review of an Additional 54 Liver Masses. Am J Clin Pathol. 2016;146(2):254–61.

34.Coopersmith CM, Lowell JA, Hassan A, Howard TK. Hepatocellular carcinoma in a patient with focal nodular hyperplasia. HPB (Oxford). 2002;4(3):135–8.

35.Haber M, Reuben A, Burrell M, Oliverio P, Salem RR, West AB. Multiple focal nodular hyperplasia of the liver associated with hemihypertrophy and vascular malformations. Gastroenterology. 1995;108(4):1256–62.

36.van Kessel CS, de Boer E, ten Kate FJ, Brosens LA, Veldhuis WB, van Leeuwen MS. Focal nodular hyperplasia: hepatobiliary enhancement patterns on gadoxetic-acid contrast-enhanced MRI. Abdom Imaging. 2013;38(3):490–501.

37.Zhou MT, Chen CS, Chen BC, Zhang QY, Andersson R. Acute lung injury and ARDS in acute pancreatitis: mechanisms and potential intervention. World J Gastroenterol. 2010;16(17):2094–9.

38.Opuszynska T. [Effect of the brand of dietary fat on rat's liver. V. Alkaline phosphatase in serum and the liver]. Rocz Panstw Zakl Hig. 1973;24(5):597–603.

39.Al-Saiedy M, Gunasekara L, Green F, Pratt R, Chiu A, Yang A, et al. Surfactant Dysfunction in ARDS and Bronchiolitis is Repaired with Cyclodextrins. Mil Med. 2018;183(suppl_1):207–15.

40.Floros J, Pavlovic J. Genetics of acute respiratory distress syndrome: challenges, approaches, surfactant proteins as candidate genes. Semin Respir Crit Care Med. 2003;24(2):161–8.

41.Lin Z, Pearson C, Chinchilli V, Pietschmann SM, Luo J, Pison U, et al. Polymorphisms of human SP-A, SP-B, and SP-D genes: association of SP-B Thr131Ile with ARDS. Clin Genet. 2000;58(3):181–91.

42.Silveyra P, Floros J. Genetic variant associations of human SP-A and SP-D with acute and chronic lung injury. Front Biosci (Landmark Ed). 2012;17:407–29.

43.Gao X, Qian P, Cen D, Hong W, Peng Q, Xue M. Synthesis of phosphatidylcholine in rats with oleic acid-induced pulmonary edema and effect of exogenous pulmonary surfactant on its De Novo synthesis. PLoS One. 2018;13(3):e0193719.

44.Hove-Jensen B, Andersen KR, Kilstrup M, Martinussen J, Switzer RL, Willemoes M. Phosphoribosyl Diphosphate (PRPP): Biosynthesis, Enzymology, Utilization, and Metabolic Significance. Microbiol Mol Biol Rev. 2017;81(1).

45.Juss JK, House D, Amour A, Begg M, Herre J, Storisteanu DM, et al. Acute Respiratory Distress Syndrome Neutrophils Have a Distinct Phenotype and Are Resistant to Phosphoinositide 3-Kinase Inhibition. Am J Respir Crit Care Med. 2016;194(8):961–73.

46.Bhargava M, Becker TL, Viken KJ, Jagtap PD, Dey S, Steinbach MS, et al. Proteomic profiles in acute respiratory distress syndrome differentiates survivors from non-survivors. PLoS One. 2014;9(10):e109713.

47.Kovach MA, Stringer KA, Bunting R, Wu X, San Mateo L, Newstead MW, et al. Microarray analysis identifies IL–1 receptor type 2 as a novel candidate biomarker in patients with acute respiratory distress

syndrome. Respir Res. 2015;16:29.

48.Evans CR, Karnovsky A, Kovach MA, Standiford TJ, Burant CF, Stringer KA. Untargeted LC-MS metabolomics of bronchoalveolar lavage fluid differentiates acute respiratory distress syndrome from health. J Proteome Res. 2014;13(2):640–9.

49.Hemnes AR, Zhao M, West J, Newman JH, Rich S, Archer SL, et al. Critical Genomic Networks and Vasoreactive Variants in Idiopathic Pulmonary Arterial Hypertension. Am J Respir Crit Care Med. 2016;194(4):464–75.

50.Korrodi-Gregorio L, Soto-Cerrato V, Vitorino R, Fardilha M, Perez-Tomas R. From Proteomic Analysis to Potential Therapeutic Targets: Functional Profile of Two Lung Cancer Cell Lines, A549 and SW900, Widely Studied in Pre-Clinical Research. PLoS One. 2016;11(11):e0165973.

51.Pediatric Acute Lung Injury Consensus Conference G. Pediatric acute respiratory distress syndrome: consensus recommendations from the Pediatric Acute Lung Injury Consensus Conference. Pediatr Crit Care Med. 2015;16(5):428–39.

52.Yehya N, Keim G, Thomas NJ. Subtypes of pediatric acute respiratory distress syndrome have different predictors of mortality. Intensive Care Med. 2018;44(8):1230–9.

53.Szilagyi KL, Liu C, Zhang X, Wang T, Fortman JD, Zhang W, et al. Epigenetic contribution of the myosin light chain kinase gene to the risk for acute respiratory distress syndrome. Transl Res. 2017;180:12–21.

54.McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297–303.

55.Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):e164.

56.Kramer A, Green J, Pollard J, Jr., Tugendreich S. Causal analysis approaches in Ingenuity Pathway Analysis. Bioinformatics. 2014;30(4):523–30.

57.Toby IT, Widmer J, Dyer DW. Divergence of protein-coding capacity and regulation in the Bacillus cereus sensu lato group. BMC Bioinformatics. 2014;15 Suppl 11:S8.

58.R Core Team. R: A language and environment for statistical computing. 3.4.0 ed. Vienna, Austria: R Foundation for Statistical Computing; 2012.

59.Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4(1):44–57.

60.Huang da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009;37(1):1–13.

61.Huang da W, Sherman BT, Zheng X, Yang J, Imamichi T, Stephens R, et al. Extracting biological meaning from large gene lists with DAVID. Curr Protoc Bioinformatics. 2009;Chapter 13:Unit 13 1.

62.Huang DW, Sherman BT, Tan Q, Collins JR, Alvord WG, Roayaei J, et al. The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. Genome Biol. 2007;8(9):R183.

Tables

Table 1: The number of variants identified from GP7 and the function of each set of variants.

SNP (Variant) Function	# of Variants (SNPs) identified from GP7
UTR3	782
Intronic or ncRNA_intronic	1343
exonic_synonymous	356
exonic_nonsynonymous	343
Intergenic	321
Upstream or downstream	144
UTR5	118
ncRNA_exonic	66
exonic_unknown	16
exonic_nonframeshift	16
exonic_frameshift	8
exonic_stopgain	2
UTR5;UTR3	1

All functions were categorized based upon the filtered list of variants derived from those highly enriched in GP7.

Table 2: Variant frequency in GP family members and unrelated ARDS cases.

[+]Family member relationship to patient	Sample identifier	Frequency of variants found in sample
Daughter	GP1	2231
Daughter	GP2	548
Husband	GP3	214
Brother	GP4	1619
Uncle (paternal)	GP5	403
Son	GP6	106
Patient	GP7	3516
Father	GP8	286
Mother	* N/A	* N/A
Unrelated ARDS case	JF9	2448
Unrelated ARDS case	JF10	950
Unrelated ARDS case	JF11	2532
Unrelated ARDS case	JF12	2747
Unrelated ARDS case	JF13	2853
Unrelated ARDS case	JF14	3008
Unrelated ARDS case	JF15	2981
Unrelated ARDS case	JF16	2044
Unrelated ARDS case	JF17	2956
Unrelated ARDS case	JF18	2103

[+]The relationship of each family member to the GP7 case is shown. Unique identifiers for family members and unrelated JF samples as well as the frequency of variants detected in each sample compared to GP7 are shown in second and third column respectively.

Table 3: IPA analysis of groups of genes, unique identifiers, and biological functions.

Analysis group ID (from Figure 4)	Number of Genes in group	Biological Description	Sample group tag
A	155	literature genes curated from ARDS-related studies	IPA/ARDS/literature genes
B	314	Exonic genes affecting stop codon and non-synonymous mutations	Exon/frameshift/stopgain/nonsyn
C	973	Exonic genes from Group B plus genes from 3' UTR, 5' UTR, and non-coding RNAs	Exon/frameshift/stopgain/nonsyn/3UTR/5UTR/ncRNAexon
D	104	Genes shared by GP7 and all JF cases	All ARDS cases shared genes

Group B, C, and D represented the genes derived from this study; Group A genes were those available in the ARDS literature for all biological pathway predictions using IPA.

Table 4. Group D genes and variants shared across GP7 and all unrelated ARDS cases.

Gene Name	Variant location and exonic information
PPT1	chr1_40539076_40539076_C_T_UTR3_
PPT1	chr1_40539203_40539203_-_TGAT_UTR3_
PPT1	chr1_40539448_40539448_A_C_UTR3_
PPT1	chr1_40548900_40548900_A_G_intronic_
SORBS1	chr10_97116219_97116219_C_T_intronic_
CYFIP1	chr15_23004124_23004127_CTAA_-_downstream_
KDSR	chr18_61022791_61022791_C_T_exonic_synonymous_SNV
MYH14*	chr19_50796905_50796905_G_A_exonic_nonsynonymous_SNV
MYH14*	chr19_50796960_50796960_C_T_intronic_
MYH14*	chr19_50813169_50813169_A_C_UTR3_
ADH5	chr4_100006645_100006645_A_G_intronic_
ADH5	chr4_100009738_100009738_G_C_intronic_
APC	chr5_112043384_112043384_T_G_UTR5_
APC	chr5_112116632_112116632_C_T_intronic_
APC	chr5_112164561_112164561_G_A_exonic_synonymous_SNV
APC	chr5_112175770_112175770_G_A_exonic_synonymous_SNV
APC	chr5_112176325_112176325_G_A_exonic_synonymous_SNV
APC	chr5_112176559_112176559_T_G_exonic_synonymous_SNV
FGFR4	chr5_176516953_176516953_A_G_intronic_
FGFR4	chr5_176517292_176517292_A_G_intronic_
FGFR4	chr5_176517326_176517326_T_C_intronic_
FGFR4	chr5_176517797_176517797_C_T_exonic_synonymous_SNV
FGFR4	chr5_176523562_176523562_C_A_intronic_
CNKSR3	chr6_154771277_154771277_A_G_exonic_nonsynonymous_SNV
MAGI2	chr7_77647322_77647322_T_C_UTR3_
MAGI2	chr7_77764591_77764591_C_A_intronic_

*Denotes the same gene represented by 3 variants which was also found present in group B genes (Table 3). All other genes and variants listed in the table were unique to ARDS cases (i.e. group D) and not found in any other group.

Table 5. Surfactant genes found in GP7 case.

Variants from Surfactant Genes	Gene_exonic function	Gene Info	CHROM	POS	POS
chr8_22020294_22020294_C_A	intronic	*SFTPC*	chr8	22020294	22020294
chr10_81316603_81316603_C_T	UTR3	*SFTPA1*	chr10	81316603	81316603
chr10_81353921_81353921_T_C	intergenic	*SFTPA2*	chr10	81353921	81353921

The column labeled “CHROM” denotes chromosomal location; the column labeled “POS” denotes the coordinate for the position of the variant within the chromosomal location.

Figures

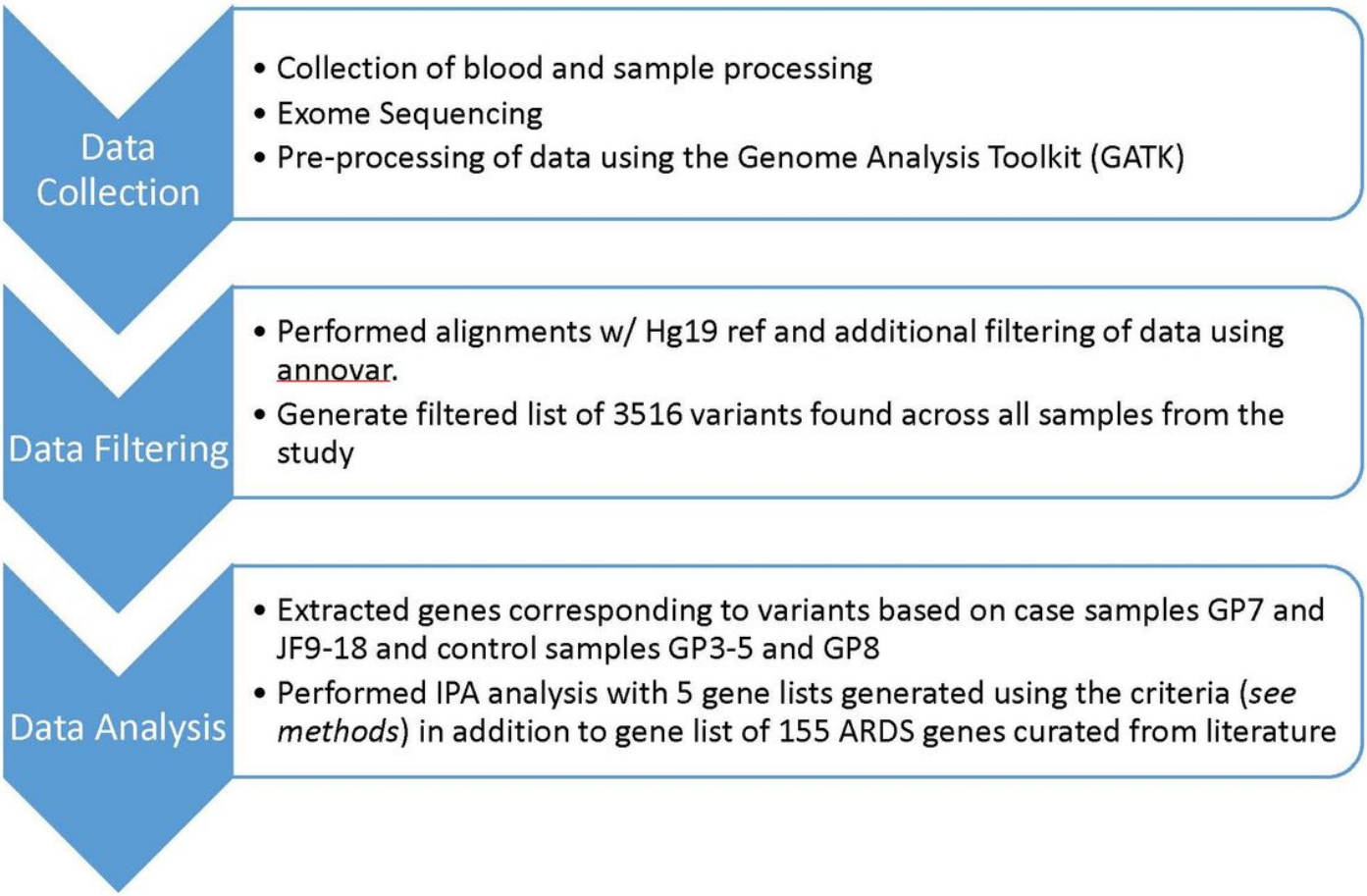

Figure 1

Analysis Workflow. Outline of the workflow steps for data collection starting from the sample processing level, data filtering steps including pre-processing of the raw sequence data file and culminating in the final data analysis steps of assessments with the post-processed data.

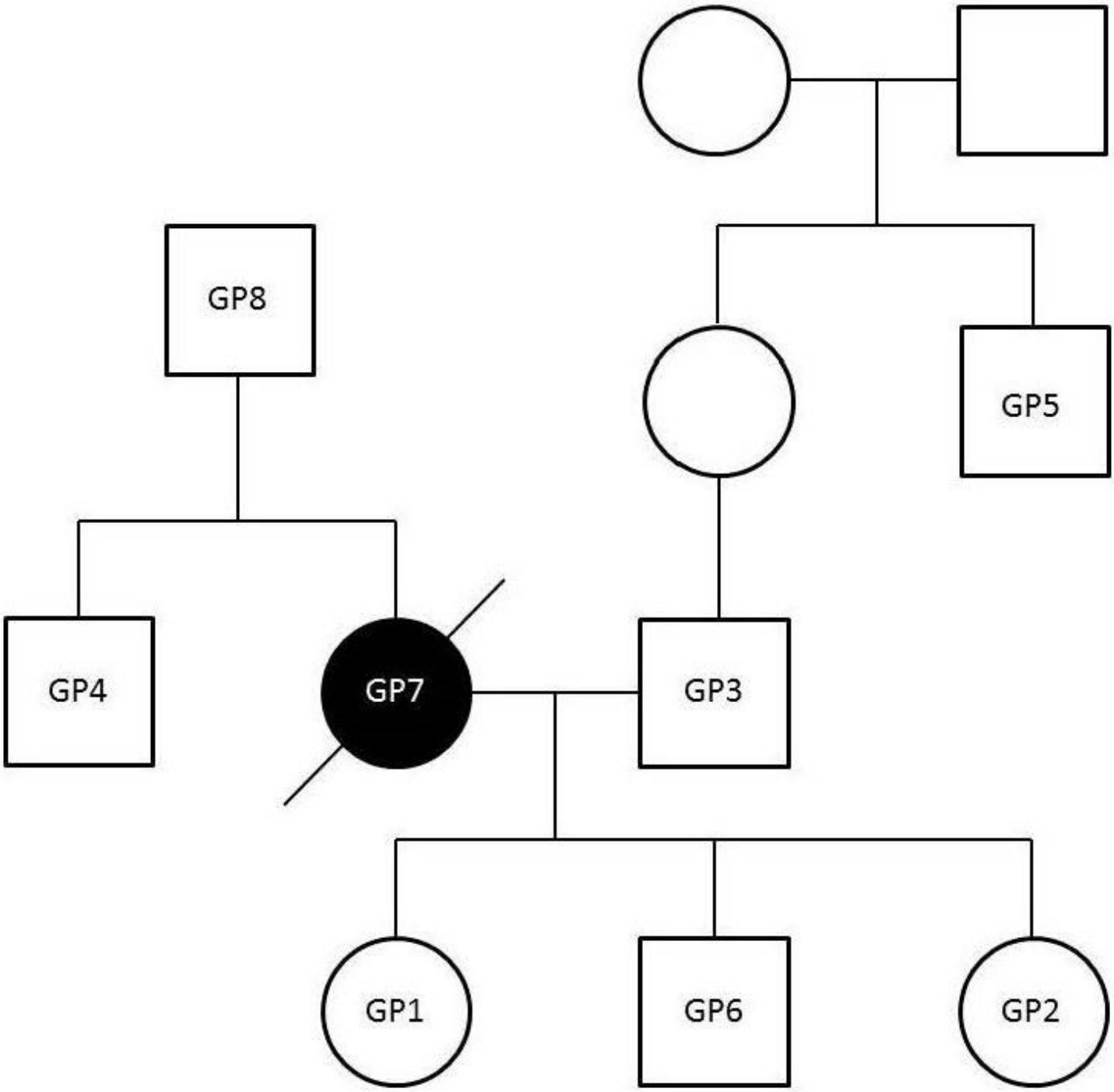

Figure 2

Family pedigree. The pedigree represents the family for our primary case, GP7. Circles denote females and squares denote males. The empty circles and square are indicating those family members whose samples were not available for the study. GP7 is a female and represents our primary case. GP4, GP8, GP6, and GP3 are males. GP1 and GP2 are females. GP1, GP2, and GP6 are the offspring of GP7 and GP3 and are also the youngest family members as they were <30 years old at the study's inception.

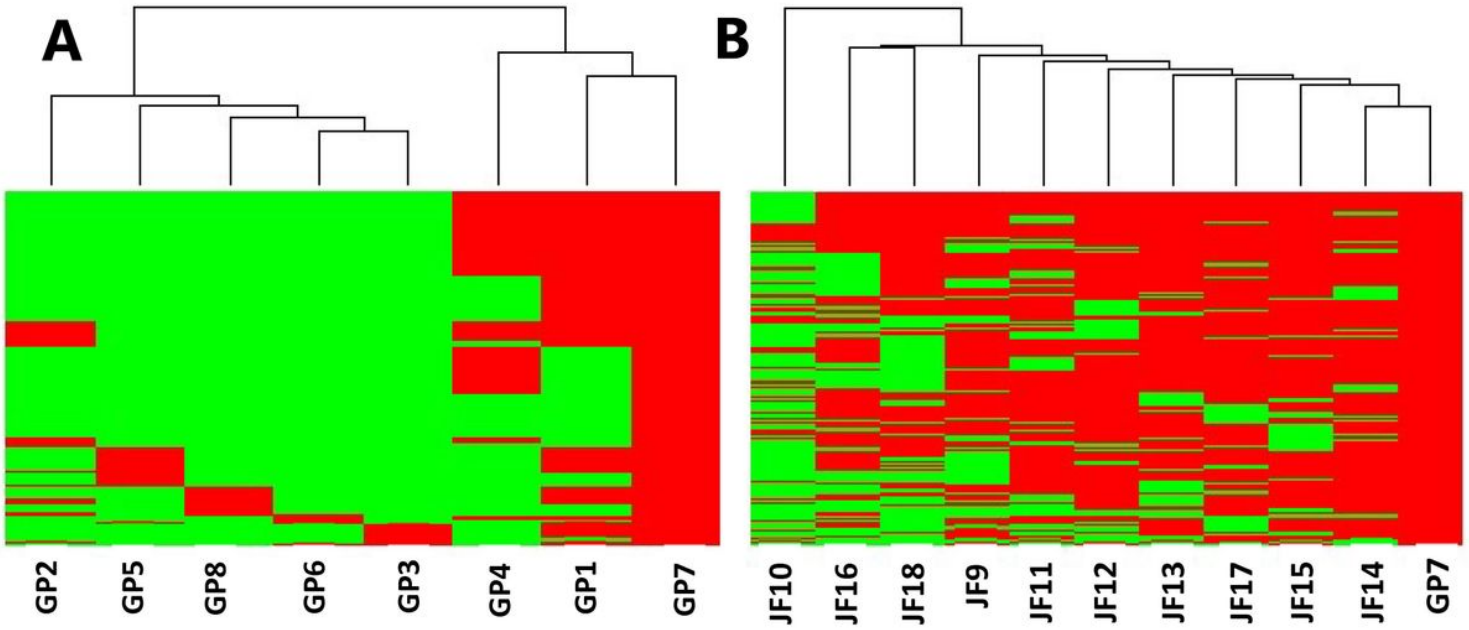

Figure 3

Heatmap of Family members and the GP7 case. The heatmap was generated using a hierarchical clustering algorithm applied to each GP sample as compared to GP7, based on the presence (colored in red) or absence (colored in green) of each individual variant. The samples clustered closest to GP7 indicate a high degree of similarity of variant occurrence. GP1 was the most similar to GP7. GP1 and GP4 form a clade with GP7. GP2 was the least similar to GP7. Heatmap of unrelated individuals with ARDS and the GP7 case. The heatmap was generated using a hierarchical clustering algorithm applied to each JF sample as compared to GP7 , based on the presence (colored in red) or absence (colored in green) of each individual variant. The samples clustered closest to GP7 indicate a high degree of similarity of variant occurrence. JF14 was the most similar sample to GP7 and JF10 was the least similar.

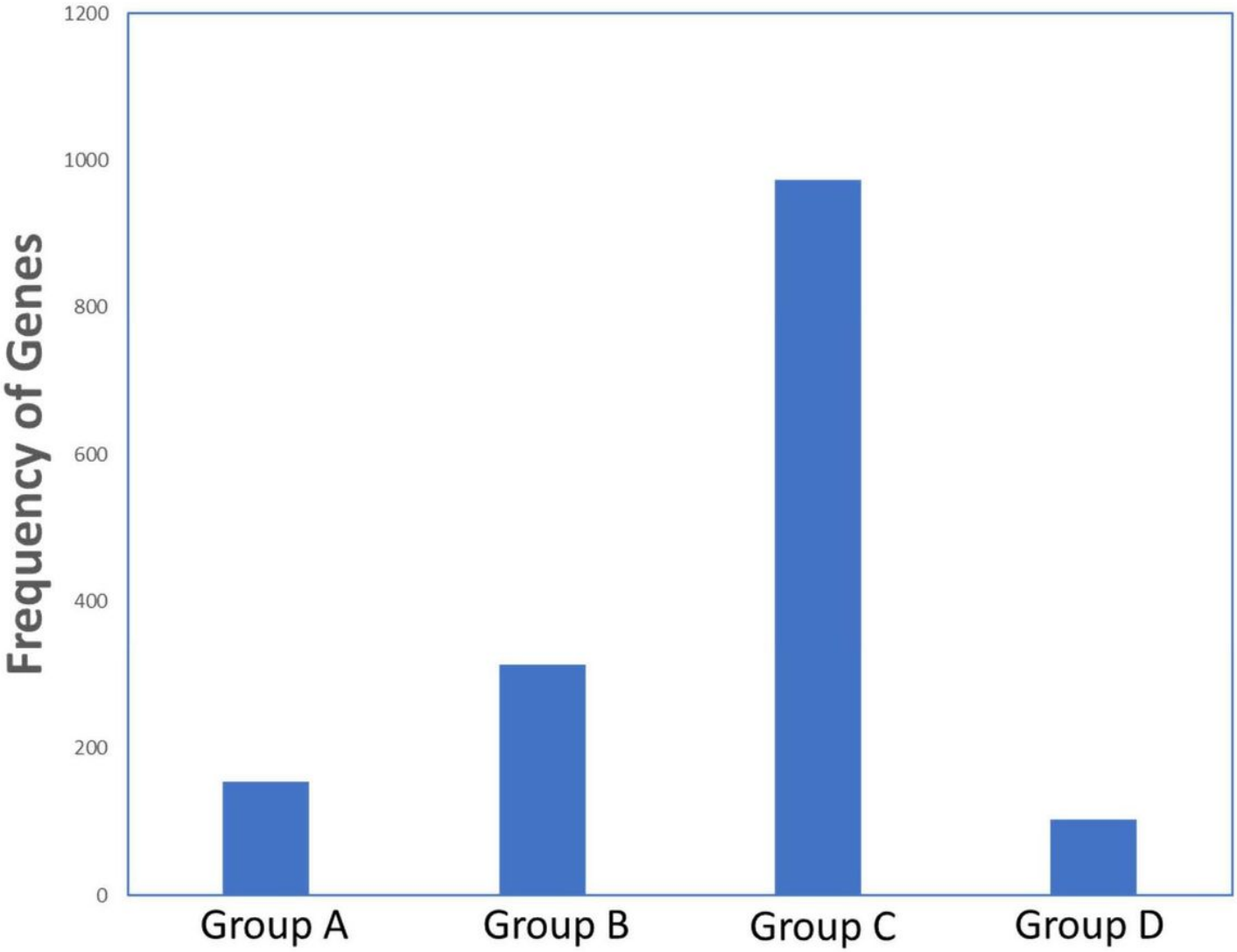

Figure 4

Relative gene frequency for IPA assessed gene lists. The variants from the study were categorized according to their respective genes and were subsequently placed in groups. They were grouped from A-D, based upon the following criteria: (1) Group A- comprised of 155 literature genes curated from ARDs-related studies; (2) Group B- comprised of 363 variants represented by 314 exonic genes affecting stop codon, genes with frameshift mutations and non-synonymous mutations; (3) Group C- comprised of 1,376 variants represented by 973 genes (which included all genes from Group B) plus genes found within the 3' UTR, 5' UTR, and non-coding RNAs and; (4) Group D- 128 variants that were represented by 104 genes shared by GP7 and all JF control cases from the study. The y-axis represents the total count of genes that belong to each group. The x-axis denotes each of the gene lists.

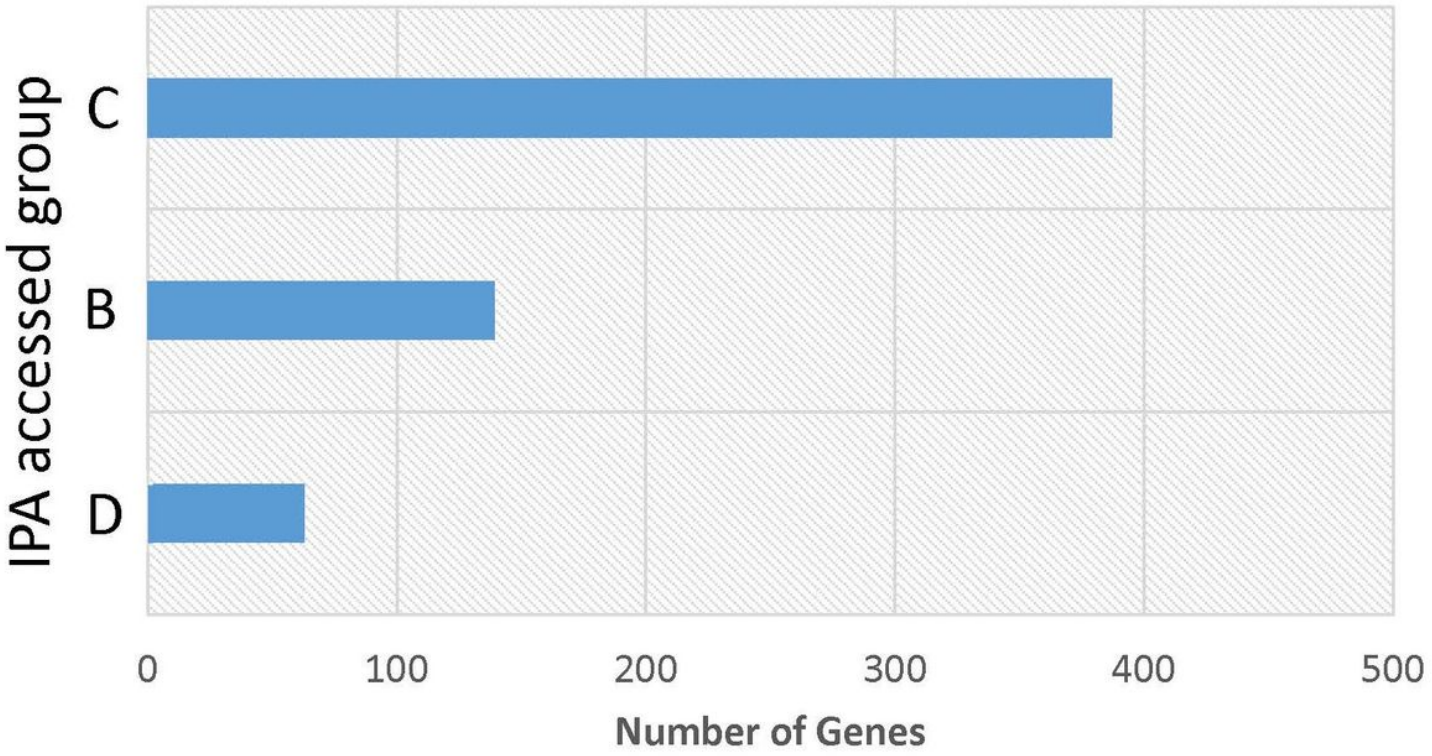

Figure 5

Novel pathway function identified from variants. Liver hyperplasia/hyperproliferation was a novel function predicted for the IPA assessed groups C, B, and D (as represented in the bar graph by C, B, and D respectively), but was not a significant pathway predicted in group A, the ARDS literature gene list. The total number of genes predicted was highest in group C, which represents all genes from group B, plus genes found within the 3' UTR, 5' UTR, and non-coding RNAs.

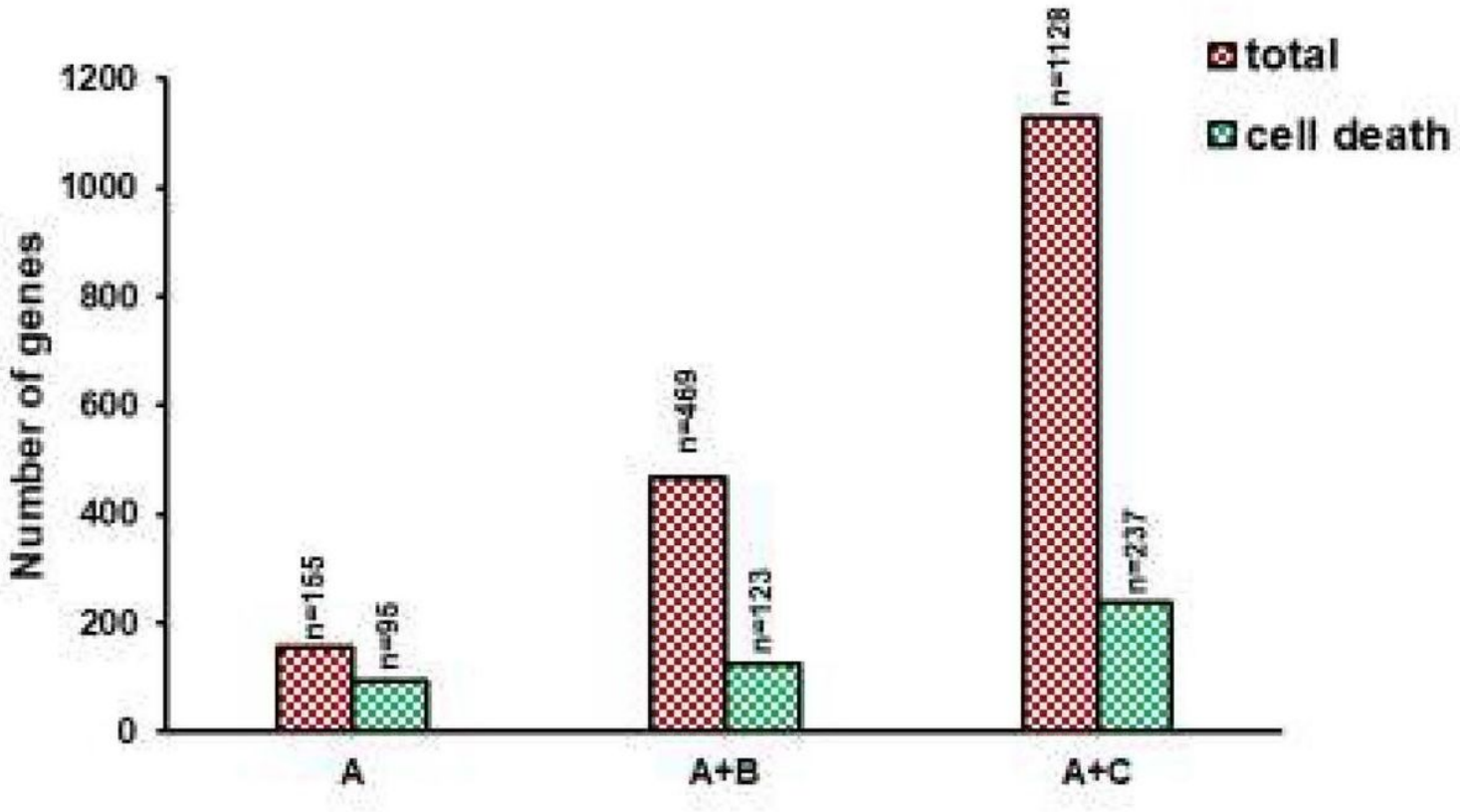

Figure 6

Genes found to be predictive of cell death and survival functions. The bar graph illustrates the number of genes for which top molecular/cellular function was cell death and survival. Assessments were done for genes from Group A (155), Group A+B (155+314), and Group A+C (155+973). Of the 95 genes from Group A, 82 genes were found to be shared amongst groups A, B and C. Forty-one additional genes were identified from Group B + Group A assessments, and 155 additional genes were identified in Group C + Group A assessments. Each pair of bars represents the frequency of genes in the search criteria and the frequency of genes identified as having a function in cell death and survival, respectively.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AdditionalFile2.xlsx
- AdditionalFile8.xlsx
- AdditionalFile3.xlsx
- AdditionalFile5.xlsx
- AdditionalFile7.xlsx
- AdditionalFile1.xlsx

- AdditionalFile6.xlsx
- AdditionalFile4.xlsx