The genomic landscape of Acute Respiratory Distress Syndrome: a meta-analysis by information content of genome-wide studies of the host response.

Jonathan E Millar[1,2], Sara Clohisey-Hendry[1,2], Megan McMannus[1], Marie Zechner[1], Bo Wang[1], Nick Parkinson[1], Melissa Jungnickel[1], Nureen Mohamad Zaki[1], Erola Pairo-Castineira[1,2], Konrad Rawlik[1,2], Clark D Russell[2], Lieuwe DJ Bos[3], Nuala J Meyer[4], Manu Shankar-Hari[2], Carolyn Calfee[5], Daniel F McAuley[6], and J Kenneth Baillie[1,2]

1. Roslin Institute, University of Edinburgh, Edinburgh, United Kingdom.
2. Centre for Inflammation Research, University of Edinburgh, Edinburgh, United Kingdom.
3. Intensive Care, Amsterdam UMC-location AMC, University of Amsterdam, Amsterdam, The Netherlands.
4. Division of Pulmonary, Allergy, and Critical Care, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania, USA.
5. Division of Pulmonary, Critical Care, Allergy & Sleep Medicine, Department of Medicine, University of California San Francisco, San Francisco, California, USA.
6. Wellcome-Wolfson Institute for Experimental Medicine, Queen's University Belfast, Belfast, United Kingdom.

Abstract

Introduction

The acute respiratory distress syndrome (ARDS) is clinically defined as acute hypoxaemic respiratory failure due to non-cardiogenic pulmonary oedema[1]. It occurs following a variety of insults; pulmonary and extra-pulmonary. While this definition has been useful in identifying patients at risk of serious morbidity and death[2], it overlooks the underlying biology and masks heterogeneity[3]. Arguably, this has contributed to limited success in developing therapeutics[4]. In contrast, a biological definition of ARDS, based on mechanistically distinct sub-phenotypes, may provide the lever necessary for future drug discovery[5].

Functional genomics technologies enable disease characterisation at unprecedented resolution. The emergence of coronavirus disease 2019 (COVID-19) has provided an opportunity to test their usefulness for drug discovery. A notable success has been the finding that baricitinib, a Janus kinase inhibitor, reduces mortality in patients hospitalised with COVID-19[6]. *A priori* support for baricitinib was greatly enhanced following the discovery of a causal link between elevated tyrosine kinase 2 (*TYK2*) expression and severe COVID-19 in genome-wide association studies (GWAS)[7,8]. The availability of omics data for non-COVID ARDS is limited by comparison, although recent studies have used these techniques to examine signatures of non-COVID ARDS sub-phenotypes[9,10].

An unresolved challenge is how large omics data can be effectively exploited[11]. Specifically, how can we combine data from heterogeneous sources to derive new insights or recalibrate our current understanding in the light of new data? We have proposed meta-analysis by information content (MAIC) as a data-driven, algorithmic, method for combining gene lists from diverse sources[12]. MAIC is agnostic to the quality or methodology of the sources and combines ranked or un-ranked gene sets by calculating weights for each list and gene, and iteratively updating them to converge on a ranked meta-list. We have successfully applied MAIC to host-genomics studies of Influenza A[12] and SARS-CoV-2[7,13], and shown that it out-performs existing algorithms when combining ranked and un-ranked lists obtained from heterogeneous sources[14].

In this work, we present a living meta-analysis by information content of ARDS host genomics studies. This serves as an open-source resource for gene prioritisation, functional genomics, and drug target discovery. An interactive interface can be accessed at https://baillielab.net/maic/ards, alongside a complementary (R package)[https://github.com/baillielab/ARDSMAICr].

Results

Systematic review

Our search yielded 8,937 unique citations (Fig. S1). We retrieved 74 articles for full-text evaluation and included 40 in our meta-analysis[9,10,15–52]. These 40 studies produced 44 unique gene lists (22 transcriptomic, 13 proteomic, and 9 based on genome-wide association studies (GWAS); see Table 1). Three studies reported results from multiple methodologies[10,34,39], and several used more than one tissue type[19,22,33]. Excluding GWAS, 14 gene lists (40%) were derived from lung or airway samples, and 21 (60%) from blood. We could not retrieve one gene list[27]. No whole-genome sequencing GWAS were found, and only 36% (n=8) of transcriptomic lists used next-generation sequencing techniques. The earliest included study was published in 2004[19], however, almost half (n=19, 47.5%) were published in the last 5 years.

Most studies aimed to identify genes or proteins associated with ARDS susceptibility (n=27, 67.5%). The remainder examined associations with survival (n=6, 15%), sub-phenotype (n=4, 10%), disease progression (n=2, 5%), or severity (n=1, 2.5%). In total, studies included 6,856 patients with ARDS. A detailed summary of study designs, demographics, and ARDS aetiology is provided in Supplementary Table 1.

Meta-analysis by information content (MAIC)

First, we analysed all 43 available gene lists using MAIC. Lists were categorised by method (i.e., GWAS, transcriptomics, and proteomics) and technique (e.g., RNA-seq, mass spectrometry; see Table 1). In total, we ranked 7,085 unique genes (or SNPs), with a median of 27 genes per list (range 1-4,954). The top 100 ranked genes are summarized in Figure 1. Most genes were found in a single category (n=5,866, 82.8%); only 157 (2.2%) were identified in ≥ 3 categories, with the maximum number of categories supporting a gene being 5 (Figure 1). Similarly, few genes (n=362, 5.1%) were identified by more than one method, with only *AKR1B10*, *HINT1*, *HSPG2*, *S100A11*, and *SLC18A1* present in transcriptomic, proteomic, and GWAS-based lists. To prioritise genes for further investigation, we used the unit invariant knee method[53] to identify the inflection point in the MAIC score curve. This prioritised 1,306 genes with scores above this point (Figure 1). These genes were more likely to be found in ≥ 2 lists or categories and by more than one method (Figure 1).

To assess the influence of individual lists, we calculated the information content (IC), reflecting the sum of gene scores across all lists (Figure 2), and the information contribution (ICtb), measuring the sum of gene scores contributing to a gene's overall MAIC score. To obtain relative values, we divided the IC/ICtb for each list by the total. This demonstrated that only 10 lists (from 9 studies) contributed >1% of the total information by either metric (Tab. S2). Notably, the RNA-seq list from Sarma et al.[10] accounts for >50% of the total IC and ICtb, a function of its length. To account for this, we normalised relative IC/ICtb by the number of genes per list. Along with the proportion of replicated genes in each list, this provides an alternative perspective, with several proteomic studies ranking highly (Figure 2).

Comparison with existing ARDS sources and COVID-19

To place our meta-analysis results in context, we evaluated the overlap between the genes prioritised by MAIC and those from two established resources: BioLitMine[54], using an ARDS MeSH search, and the ARDS Database of Genes[55] (Fig. S2a and Fig. S2c). BioLitMine identified 271 ARDS-associated genes, of which 142 (52.4%) were present in our analysis. Almost half of the overlapping genes (n = 63, 44.4%) ranked within our prioritised set (Tab. S3). Of the 239 genes catalogued in the ARDS Database of Genes, 177 (74.1%) were also found in our study. However, both sources contain gene associations lacking genome-wide support.

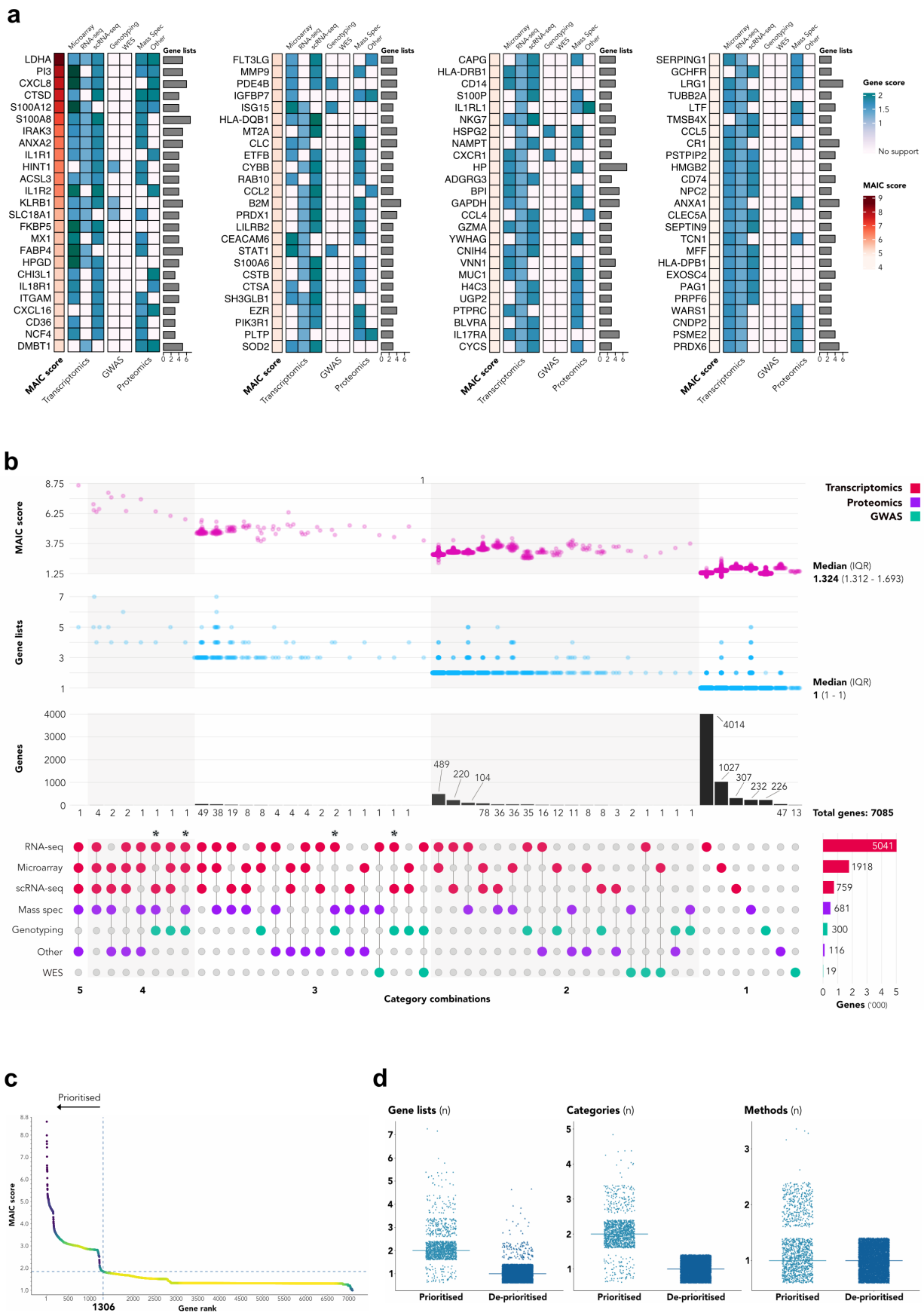

Figure 1: **Meta-analysis by information content**. (a) Heatmap of top 100 ranked genes showing MAIC score, highest score per category, and number of supporting lists. (b) UpSet plot of MAIC genes showing total numbers for each category combination, MAIC score distribution, and supporting lists. (b) Gene prioritization using the Unit Invariant Knee method. Intersection of lines identifies elbow point of best-fit curve. 1,306 genes in upper left quadrant were prioritied. (c) Strip plots comparing number of lists and categories/methods per gene between prioritized and deprioritized sets.

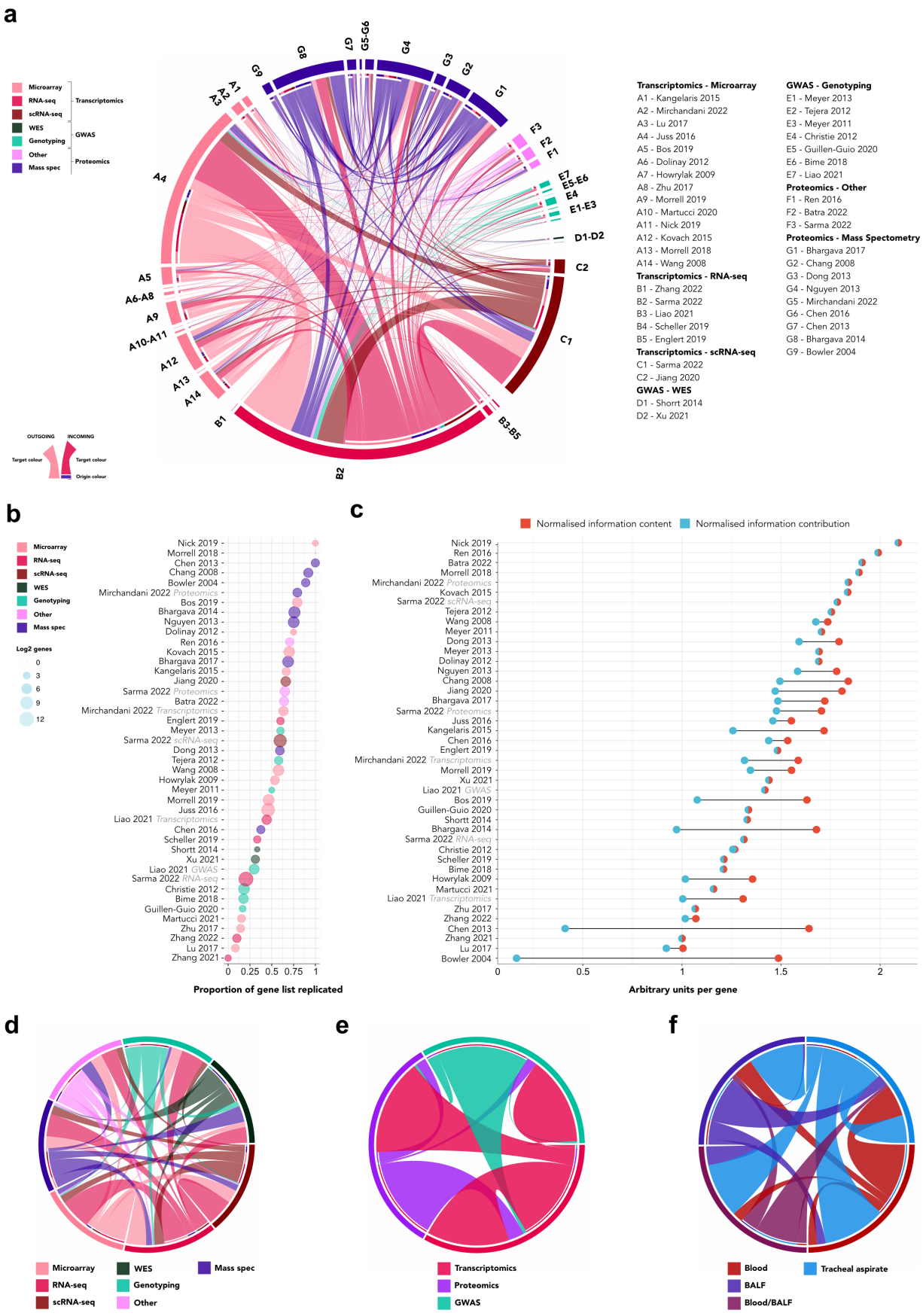

Figure 2: **Attributing information in MAIC**. (a) Shared information content (IC) between gene lists. Links indicate absolute IC (sum of common gene scores) between studies. (b) Proportion of replicated genes. Circle diameter is logarithm (base 2) of gene number per list. (c) IC normalized by number of genes. Overlapping circles denote equal normalized IC and contribution (ICtb - sum of common gene scores contributing to MAIC), indicating all gene scores contributed to MAIC. (d) Shared IC between categories, scaled so links show fraction of total IC. (e) Shared IC between methods, scaled. (f) Shared IC between tissue types, scaled.

Table 1: Summary of studies and gene lists included in the systematic review

Year	Study	Focus	Definition	N[a]	Method	Technique	Tissue	Cell type
2022	Batra[15]	Survival	Berlin	24	Proteomics	Other	Blood	
	Mirchandani[39]	Susceptibility	Berlin	22	Proteomics	Mass Spec	Blood	Monocytes
					Transcriptomics	Microarray	Blood	Monocytes
	Sarma[10]	Sub-phenotype	Berlin	41	Proteomics	Other	TA	
					Transcriptomics	RNA-seq	TA	
					Transcriptomics	scRNA-Seq	TA	Immune cells
	Zhang[51]	Susceptibility	AECC	11	Transcriptomics	RNA-Seq	Blood	Exosomes
2021	Liao[34]	Survival	Either	390	GWAS	Genotyping	Blood	
					Transcriptomics	RNA-seq	Blood	PBMCs
	Martucci[36]	Sub-phenotype	None	11	Transcriptomics	Microarray	Blood	
	Xu[49]	Survival	Berlin	105	GWAS	WES	Blood	
	Zhang[50]	Susceptibility	Berlin	5	Transcriptomics	RNA-seq	Blood	
2020	Guillen-Guio[28]	Susceptibility	Berlin	633	GWAS	Genotyping	Blood	
	Jiang[30]	Susceptibility	Berlin	3	Transcriptomics	scRNA-seq	Blood	PBMCs
2019	Bos[9]	Sub-phenotype	Berlin	210	Transcriptomics	Microarray	Blood	
	Englert[26]	Susceptibility	Either	11	Transcriptomics	RNA-seq	Blood	
	Morrell[41]	Survival	AECC	36	Transcriptomics	Microarray	BALF	AMs
	Scheller[45]	Susceptibility	None	6	Transcriptomics	RNA-seq	BALF	EVs
2018	Bime[18]	Susceptibility	Either	232	GWAS	Genotyping	Blood	
	Morrell[40]	Susceptibility	Berlin	35	Transcriptomics	Microarray	BALF	
2017	Bhargava[17]	Survival	AECC	36	Proteomics	Mass Spec	BALF	
	Lu[35]	Susceptibility	AECC	12	Transcriptomics	Microarray	Blood	
	Zhu[52]	Susceptibility	Berlin	199	Transcriptomics	Microarray	Blood	

Year	Study	Focus	Definition	N[a]	Method	Technique	Tissue	Cell type
2016	Chen[22]	Severity	AECC	7	Proteomics	Mass Spec	BALF/Blood	
	Juss[31]	Susceptibility	Berlin	23	Transcriptomics	Microarray	Blood	Neutrophils
	Nick[42]	Sub-phenotype	AECC	121	Transcriptomics	Microarray	Blood	Neutrophils
	Ren[44]	Susceptibility	Berlin	14	Proteomics	Other	Blood	
2015	Kangelaris[32]	Susceptibility	Berlin	29	Transcriptomics	Microarray	Blood	
	Kovach[33]	Susceptibility	AECC	18	Transcriptomics	Microarray	BALF/Blood	AMs
2014	Bhargava[16]	Progression	AECC	22	Proteomics	Mass Spec	BALF	
	Shortt[46]	Susceptibility	AECC	213	GWAS	WES	Blood	
2013	Chen[21]	Susceptibility	Berlin	11	Proteomics	Mass Spec	Blood	
	Dong[25]	Progression	None	14	Proteomics	Mass Spec	BALF	AMs
	Meyer[38]	Susceptibility	Berlin	661	GWAS	Genotyping	Blood	
	Nguyen[43]	Susceptibility	AECC	30	Proteomics	Mass Spec	BALF	
2012	Christie[23]	Susceptibility	AECC	812	GWAS	Genotyping	Blood	
	Dolinay[24]	Susceptibility	AECC	35	Transcriptomics	Microarray	Blood	
	Tejera[48]	Susceptibility	AECC	1400	GWAS	Genotyping	Blood	
2011	Frenzel[27]	Survival	AECC	46	Proteomics	Mass Spec	BALF	
	Meyer[37]	Susceptibility	AECC	1241	GWAS	Genotyping	Blood	
2009	Howrylak[29]	Susceptibility	AECC	13	Transcriptomics	Microarray	Blood	
2008	Chang[20]	Susceptibility	None	20	Proteomics	Mass Spec	BALF	
	Wang[47]	Susceptibility	AECC	8	Transcriptomics	Microarray	Blood	
2004	Bowler[19]	Susceptibility	AECC	16	Proteomics	Mass Spec	BALF/Blood	

a - The number of patients with ARDS included in each study. Abbreviations: AECC - American-European Consensus Conference; AMs - Alveolar macrophages; BALF - Bronchoalveolar lavage fluid; EVs - Extracellular vesicles; GWAS - Genome-wide association study; MS - Mass spectometry; PBMCs - Peripheral blood mononuclear cells; TA - Tracheal aspirate; WES - Whole-exome sequencing.

After correcting historical gene symbol aliases, we matched 4 additional genes from the BioLitMine search not initially found in the ARDS MAIC set. A further 104 genes from this search were supported by just a single publication (Fig S2b). For each of the remaining 21 genes, we obtained the 100 most co-expressed genes using ARCHS4[56] (returning data for 18) and assessed the overlap of these sets with ARDS MAIC; two-thirds exhibited <50% overlap (Fig. S2b). Finally, we compared the overlap between genes ranked by ARDS MAIC and those identified in a previous MAIC of the host response to COVID-19[13] (Fig. S2d). In total, 2,606 ARDS-associated genes (36.8%) were common to both analyses, of which 143 were prioritised by both (Fig. S2e).

Tissue and cell-specific expression

Although most gene lists were derived from blood sampling, most genes were identified in airway samples (n=5,847, 82.5%) (Fig. S3a). This was also true for the prioritized gene set, however most of these were also identified in blood (n=818, 62.6%) (Fig. S3b). For the genes uniquely identified in lists from blood samples (n=1,238), almost three-quarters are known to be expressed in the lung (HPA scRNA-seq data, ≥ 5 normalised transcripts per million (nTPM)), with a quarter being highly-expressed (≥ 100 nTPM) (Fig. S3c). For prioritized lung genes, there is a wide variety of cell-specific expression (Fig. S3d). However, in the smaller set of prioritized genes identified only in blood, clusters of expression specific to neutrophils, T cells, and monocytes are evident (Fig. S3e). Cell-type specific gene enrichment analysis suggests innate immune as well as epithelial and endothelial cell types are enriched among genes identified in airway samples (Fig. S3f). However, enrichment of epithelial and endothelial cells is not evident for prioritized genes identified from blood sampling alone (Fig. S3g).

Functional enrichment

Having identified a set of prioritised genes, we undertook several functional enrichment analyses. First, we performed over-representation analysis (ORA). In Reactome, 51 terms were significantly enriched ($P < 0.001$) (Figure 3). As expected, neutrophil degranulation and several innate immune pathways (e.g., IL-10 signalling, interferon signalling, MHC II antigen presentation, TLR4 cascade) feature heavily. However, multiple pathways associated with cholesterol biology and metabolism (e.g., chylomicron assembly/remodelling, GLUT4 translocation, TP53 regulation of metabolic genes, insulin regulation) were also over-represented. Similarly, lipid and cholesterol metabolism, as well as hyper-lipidaemia, were over-represented in KEGG and WikiPathways (Fig. S4a and Fig. S4b). In an enrichment analysis using the GWAS Catolog, the prioritised set of genes was associated with asthma (adult onset/time to onset), monocyte, lymphocyte, and eosinophil counts, aspartate aminotransferase levels, and levels of apolipoprotein A1 (Fig. S4d).

Next, we used the prioritised set of genes to create a protein-protein interaction (PPI) network. We graph-clustered this network, identifying 48 clusters with ≥ 5 members. Among the 10 largest clusters, we found programs associated with the proteaosome, cholesterol metabolism, interferon signalling, IL-6 signalling, and the complement cascade (Fig. S5). We then sought to use the PPI network to identify hub genes using an ensemble of topological methods. This analysis suggested 51 genes central to the wider network (Figure 3). Clustering these genes alone identified 5 clusters, which may be associated with innate immune cytokine signalling, interferon signalling, MHC class II antigen presentation, PI3K-Akt signalling, and eukaryotic translation elongation (Figure 3). The majority of hub genes (n=31, 61%) are currently druggable and include targets such as *IL-6*, *IL-17A*, *IL-18*, and *MAP3K14*.

Sub-groups

A source of tension in our approach is the balance between dispartiy in study designs (e.g., susceptibilty, sub-phenotype) and the requirement for sufficient data to make meaningful inferences. To address this, we undertook MAIC on sub-

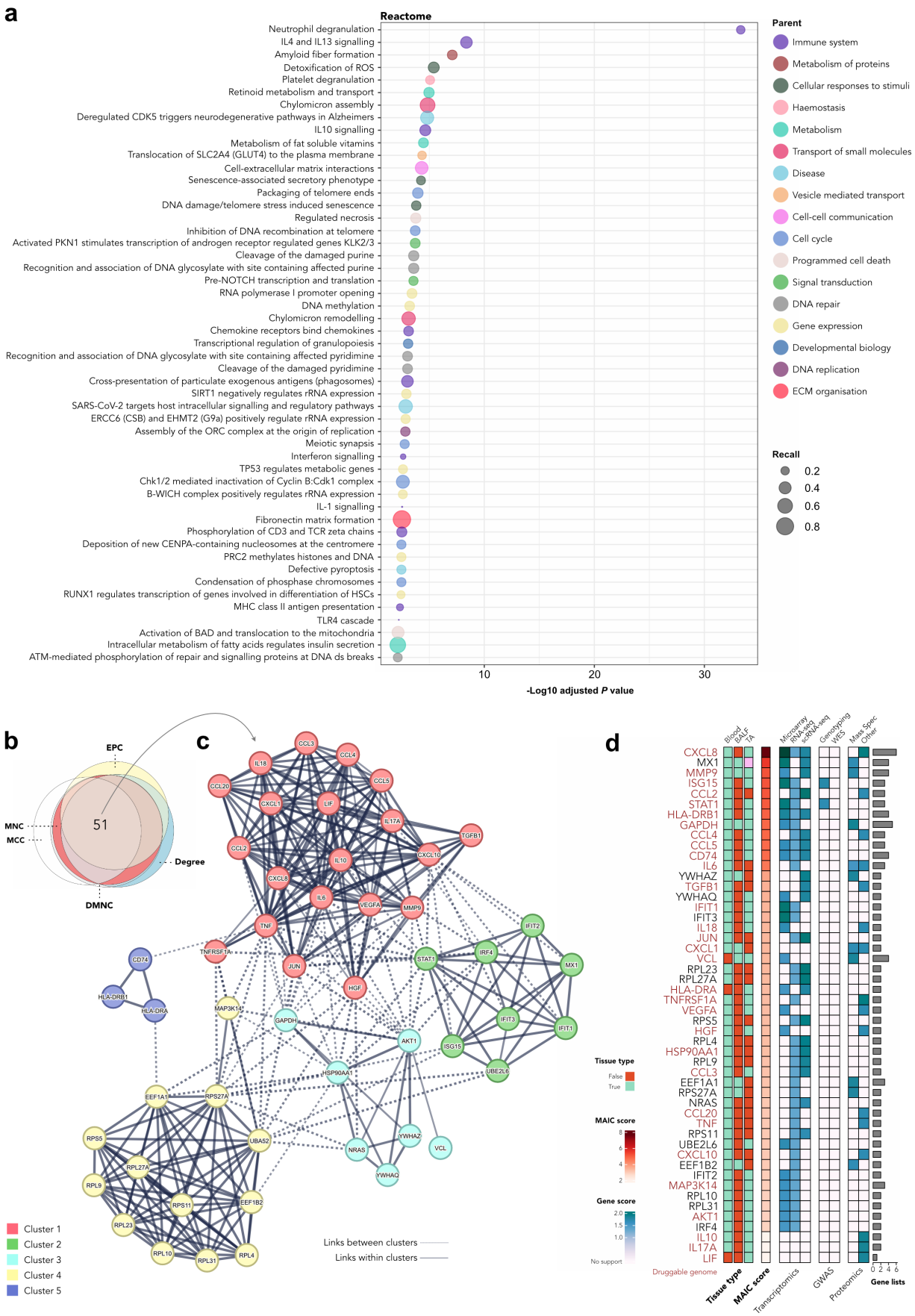

Figure 3: **Functional enrichment of prioritised genes**. (a) Significantly enriched Reactome terms ($P < 0.01$). Terms colored by parent class and size proportional to recall. (b) Euler diagram of the overlap of hub genes identified by five methods. MNC - Maximum Neighbourhood Component, MCC - Maximal Clique Centrality, DMNC - Density of MNC, EPC - Edge Percolated Component. (c) Protein-protein interaction (PPI) network of hub genes, clustered using the Markov Chain Algorithm. (d). Heatmap of common hub genes displaying tissue type(s), MAIC score, highest category score, supporting lists, and presence in the druggable genome.

sets of gene lists, stratified by study focus. There were sufficient lists to make this tractable for studies focused on susceptibility to ARDS (n=28) and studies of ARDS survival and severity (n=7) (Figure 4).

For susceptibility, there were 15 transcriptomic (54%), 7 GWAS (25%), and 6 proteomic lists (21%). MAIC ranked 2,096 genes (Figure 4). The majority of these (n=1,222, 58%) were unique to susceptibility-based lists (Figure 4). Most were identified in blood, with a small fraction found solely in airways samples. The inflection point method prioritised the top ranked 130 genes (Fig. S6a). In comparison to the BioLitMine search and the ARDS Database of Genes, 71/271 and 117/239 genes were found among the ARDS MAIC susceptibility set respectively (Fig. S6b). A microarray-based transcriptomic list from Juss *et. al.*[31] accounted for more than half (54.7%) of the relative ICtb, with an additional 12 lists having a relative ICtb ≥ 1% (Tab. S4). ORA using Reactome, KEGG, and WikiPathways identified 25 significantly enriched pathways including multiple terms related to cholesterol metabolism and glycolysis (Fig. S7a). A consensus of topological models identified 7 hub genes within a PPI network of prioritised genes. These genes cluster in a single group, characterised as being related to cholesterol metabolism by several pathway databases (Figure 4).

For survival, the 8 gene lists consisted of 3 transcriptomic lists (37.5%), 3 proteomic lists (37.5%), and 2 GWAS (25%). MAIC ranked 463 genes (Figure 4). Approximately half of these (n=238, 51%) were unique to survival-based lists. In contrast to the susceptibility analysis, most survival genes were found in airways samples. Thirty-three genes were prioritised (Fig. S6d). In total, 32/271 of the BioLitMine ARDS-associated genes and 23/239 of the ARDS Database of Genes genes were found among the ARDS MAIC survival set (Fig. S6e). The proteomic and transcriptomic lists from Bhargava *et.al*[17] and Morrell *et. al*[41] each contributed approximately 30% of the relative ICtb (Tab. S5). IL-10 and IL-18 signalling pathways were both significantly enriched in ORA (Fig. S7c). Graph-based clustering of the prioritised set of survival genes identified a single large cluster of immune-related genes including, *IL-10*, *CXCL8*, *TNFRSF1A*, and *IL2RA* (Fig. S7d).

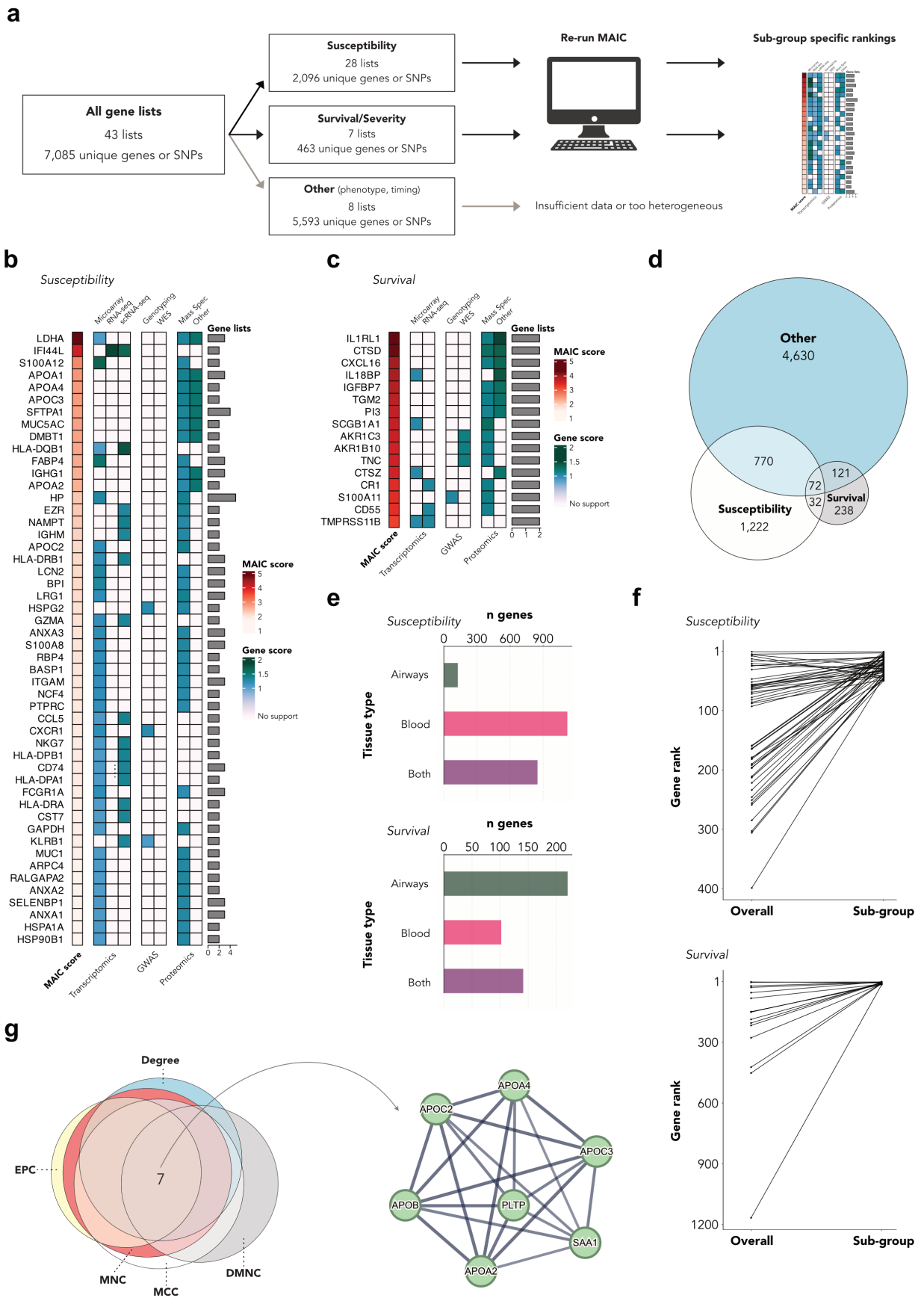

Figure 4: **MAIC of sub-groups**. (a) Schematic of ARDS MAIC sub-group analyses. (b) Heatmap of top 50 ranked genes in the susceptibility set showing MAIC score, highest score per category, and number of supporting lists. (c) Heatmap of 16 ranked genes in the survival set with multi-list support showing MAIC score, highest score per category, and number of supporting lists. (d) Eular diagram of gene overlap between the susceptibility and survival sets and the remainder of genes. (e) Bar plots of the tissue type in which genes are identified. (f) plot comparing the ranks of susceptibility and survival prioritised genes with their ranks in the full iteration of ARDS MAIC. (g) Euler diagram of the overlap of hub genes identified by five methods. MNC - Maximum Neighbourhood Component, MCC - Maximal Clique Centrality, DMNC - Density of MNC, EPC - Edge Percolated Component and a protein-protein interaction (PPI) network of hub genes, clustered using the Markov Chain Algorithm - for susceptibility.

Discussion

Our systematic integration and re-analysis of more than 20 years of omics data is the first large-scale meta-analysis of the genomic landscape of ARDS. This implicates and ranks over 7,000 genes and prioritises 1,306. The wide inclusion criteria capture a diverse range of study designs and methods and the strength of MAIC is its ability to establish the sum of this knowledge, while downgrading noisy or irrelevant information. These results have three main applications. First, they can be used to better understand the pathobiology of ARDS, providing a resource to prioritise future *in-vitro* and *in-vivo* studies and permitting comparisons between important sub-groups. Second, they prioritise therapeutic targets, serving as a source against which novel and repurposed drugs treatments can be screened. Third, they serve as a base for identifying the novelty or additive nature of future omics studies in ARDS.

Our approach has limitations. We purposefully sought studies with genome-wide hypotheses, excluding single-gene or candidate genetics studies. In the case of a gene with extensive evidence from the latter, our methodology may underestimate its association with ARDS. However, these study designs are subject to other biases, such as publication bias and investigator-driven hypotheses. In our iteration of MAIC, we did not account for direction of expression or effect. For a given gene, if the direction of expression differs between studies, MAIC may overestimate the strength of evidence associated with that gene. The inability to account for directionality also limits the scope of functional enrichment analyses which can be performed. Similarly, the use of an unsupervised prioritisation threshold may influence the outcomes of downstream analysis, however principled the approach. Finally, the general paucity of data, and in particular the limited number of single-cell transcriptomics (or proteomics) studies, limits the depth of inference that can be made. It is likely that many pathological perturbations occur with cell-specificity, which may not be apparent in a largely bulk analysis of heterogeneous tissues. The future addition of data from these modalities may reveal precision targets.

Crucially, we provide an open platform and tools to enable deeper mining of the output. This may allow others to re-analyse the data based on alternative sub-group divisions or to integrate unseen information. Ongoing multi-omic data integration with this initial study will further enhance the resolution of the data and increase our confidence in the results.

In summary, by systematically integrating decades of ARDS genomics, this study enhances are scope for gene prioritisation and enhances molecular pathophysiological clarity. Our results strongly implicate endothelial dysfunction and cholesterol metabolism dysregulation, providing a specific therapeutic targets. Enrichment patterns and sub-group differences also give clues to genomic drivers of susceptibility, outcomes, and mortality. This substantially improves our conceptualisation of the genomic landscape of ARDS, setting the stage for functional validation.

Methods

The systematic review and meta-analysis protocol was registered with the International Prospective Register of Systematic Reviews (PROSPERO; CRD42022306270). The review is reported in compliance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines[57].

Search strategy and selection criteria

A detailed description of our search strategy and eligibility criteria is provided in the Supplementary Methods. Briefly, we searched MEDLINE, Embase, bioRxiv, medRxiv, the ARDS Database of Genes[55], and the NCBI Gene Expression Omnibus from inception to April 1st, 2023 without language restrictions. We also performed single-level backwards and forwards citation searches using SpiderCite[58] and hand-searched recent review articles[59–62].

We included human genome-wide studies reporting associations between genes, transcripts, or proteins and ARDS susceptibility, severity, survival, or phenotype, accepting any contemporaneous ARDS definition. We excluded paediatric studies (age < 18 years), animal studies, *in-vitro* human ARDS models, candidate *in-vivo* or *in-vitro studies* (< 50 genes/proteins), candidate gene associations, and studies with < 5 patients per arm (except scRNA-seq).

Outcomes

We retrieved ranked lists of genes associated with the ARDS host response, preferring measures of significance and adjusted *P* values over raw *P* values when multiple ranking measures were used. We obtained both summary lists (all implicated genes) and author-defined subgroup lists. To combine subgroup lists into summary lists, we took the minimum *P* value or maximum effect size. We excluded genes below the author-defined threshold for significance/effect magnitude. If unavailable, we excluded genes with $P > 0.05$, z-score < 1.96, or log fold change < 1.5.

Study selection and data extraction

Article titles and abstracts from our search were stored in Zotero v6.0-beta (Corporation for Digital Scholarship, United States). Titles were initially screened by one author using Screenatron[58]. Two authors then independently screened abstracts against eligibility criteria, with a third resolving inconsistencies. Full texts and supplements of eligible studies were retrieved and inclusion adjudicated by consensus.

Data were extracted by one author and cross-checked by a second. Gene, transcript, or protein identifiers were mapped to HGNC symbols or Ensembl/RefSeq equivalents if no HGNC symbol was available. Unannotated SNPs were searched in NCBI dbSNP. miRBase (University of Manchester, United Kingdom) provided miRNA symbols. For microarray probes without symbols, we used the DAVID Gene Accession Conversion tool (Laboratory of Human Retrovirology and Immunoinformatics, Frederick National Laboratory for Cancer Research, United States) to map them to HGNC symbols. We extracted information relating to study design, methodology, tissue/cell type, demographics, ARDS aetiology, risk factors, severity, and outcomes.

Meta-analysis by information content (MAIC)

The MAIC algorithm has been described in detail[7,12–14]. Full documentation and the source code are available at https://github.com/baillielab/maic. Briefly, MAIC combines ranked and unranked lists of related named entities, such as genes, from heterogeneous experimental categories, without prior regard to the quality of each source. The algorithm makes four key assumptions; (1) genes associated with ARDS exist as true positives, (2) a gene is more likely to be a true positive if it is found in more than one source, (3) the probability of being a true positive is enhanced if the gene

appears in a list that contains a higher proportion of replicated genes, and (4) the probability is further enhanced if it is found in more than one category of experiment. Based on these assumptions, MAIC compares lists with each other, forming a weighting for each source based on its information content, which is then used to calculate a score for each gene. The output is a ranked list summarizing the total information supporting each gene's association with ARDS. We have shown MAIC outperforms available algorithms, especially with ranked and unranked heterogeneous data[14].

As our primary analysis, we performed MAIC on all summary gene lists, regardless of study focus. Lists were assigned categories based on their methodology and experimental technique: genome-wide association study (GWAS) - genotyping, GWAS - whole exome sequencing, transcriptomics - microarray, transcriptomics - RNA-sequencing (RNA-seq), transcriptomics - single cell RNA-seq (scRNA-seq), proteomics - mass spectometry, and proteomics - other. For secondary analyses, we performed MAIC on subsets of lists based on study focus (i.e., susceptibility to ARDS or survival/severity).

In secondary analyses, we repeated this pipeline for gene lists arising from studies in which the focus was susceptibility to ARDS or ARDS survival/severity.

For each MAIC iteration, we prioritised genes with sufficient evidentiary support for further study (i.e., the gene set before which information content diminished such that there was little/no corroboration for the remainder's ARDS association). We used the unit invariant knee method[53,63] to identify the elbow point in the best-fit curve of MAIC scores. Genes with values above this point were prioritized for downstream analyses.

ARDS literature and SARS-CoV-2 associations

We used BioLitMine[54] to query the NCBI Gene database for genes associated with the Medical Subject Heading (MeSH) term "Respiratory Distress Syndrome, Acute", generating a list of genes and publications. We descriptively compared the overlap between this list and the MAIC-ranked gene list. Similar comparisons were made between the ARDS MAIC results and the gene set in the ARDS Database of Genes[55] and a prior MAIC of SARS-CoV-2 host genomics[13].

Tissue expression and enrichment

Transcript and protein expression data for genes included in ARDS MAIC were retrieved from the Human Protein Atlas (HPA, version 21.0)[64]. We investigated mRNA expression in a consensus scRNA-seq dataset of 81 cells from 31 sources (https://www.proteinatlas.org/about/assays+annotation#singlecell_rna) and in the HPA RNA-seq blood dataset[65], containing expression levels in 18 immune cell types and total peripheral blood mononuclear cells. To investigate protein expression, we retrieved tissue-specific expression scores from the HPA[66]. We conducted cell-type specific enrichment analysis using WebCSEA[67] and extracted the top 20 general cell types for each query.

Functional enrichment

We performed functional enrichment of genes against the universe of all annotated genes using g:Profiler[68]. The following data sources were used; Kyoto Encyclopaedia of Genes and Genomes (KEGG)[69], Reactome[70], WikiPathways[71], and Gene Ontology[72]. Multiple testing was corrected for using the g:SCS algorithm[68], with a threshold of $P < 0.01$. Input lists were ordered by MAIC score were appropriate. In the case of GO cellular component terms, we used the REVIGO tool to perform multi-dimensional scaling of the matrix of all pairwise semantic similarities[73]. Enrichment was also performed against the National Human Genome Research Institute GWAS Catalog[74] using the Enrichr web-interface[75]. Protein-protein interaction enrichment was performed using STRING v11[76]. We included all possible

interaction sources but specified a minimum interaction score of 0.7. We used the the whole annotated genome as the statistical background. Markov Clustering Analysis (MCL) was applied to the resulting network with an inflation parameter of 3. Clusters were annotated by hand having considered enrichment against KEGG, Reactome, and WikiPathways. To identify hub genes within the PPI network, we used cytoHubba[77] and Cytoscape[78]. The highest ranked genes by Maximum Neighbourhood Component (MNC), Maximal Clique Centrality (MCC), Density of MNC (DMNC), Edge Percolated Component (EPC), and node degree were retrieved. The intersecting genes of these methods were deemed hub genes. Hub genes were searched for in the Drug Gene Interaction Database[79] to identify if they were present in the druggable genome.

Software and code availability

MAIC is implemented in Python v3.9.7 (Python Software Foundation, Wilmington, United States). All other analyses were performed with R v4.2.2 (R Core Team, R Foundation for Statistical Computing, Vienna, Austria). Code required to reproduce the analyses is available at https://github.com/JonathanEMillar/ards_maic_analysis. An R package (ARDSMAICR) containing the data used in this manuscript and several functions helpful in analyses is available at https://github.com/baillielab/ARDSMAICr.

References

1. ARDS Definition Task Force *et al.* Acute respiratory distress syndrome: The berlin definition. *JAMA* **307**, 2526–2533 (2012).
2. Bellani, G. *et al.* Epidemiology, patterns of care, and mortality for patients with acute respiratory distress syndrome in intensive care units in 50 countries. *JAMA* **315**, 788–800 (2016).
3. Wilson, J. G. & Calfee, C. S. ARDS subphenotypes: Understanding a heterogeneous syndrome. *Crit. Care* **24**, 102 (2020).
4. Laffey, J. G. & Kavanagh, B. P. Negative trials in critical care: Why most research is probably wrong. *Lancet Respir. Med.* **6**, 659–660 (2018).
5. Bos, L. D. J. *et al.* Towards a biological definition of ARDS: Are treatable traits the solution? *Intensive Care Med. Exp.* **10**, 8 (2022).
6. Peter W Horby, and *et al.* Baricitinib in patients admitted to hospital with COVID-19 (RECOVERY): A randomised, controlled, open-label, platform trial and updated meta-analysis. (2022) doi:10.1101/2022.03.02.22271623.
7. Pairo-Castineira, E. *et al.* Genetic mechanisms of critical illness in COVID-19. *Nature* **591**, 92–98 (2021).
8. Kousathanas, A. *et al.* Whole-genome sequencing reveals host factors underlying critical COVID-19. *Nature* **607**, 97–103 (2022).
9. Bos, L. D. J. *et al.* Understanding heterogeneity in biologic phenotypes of acute respiratory distress syndrome by leukocyte expression profiles. *Am. J. Respir. Crit. Care Med.* **200**, 42–50 (2019).
10. Sarma, A. *et al.* Hyperinflammatory ARDS is characterized by interferon-stimulated gene expression, t-cell activation, and an altered metatranscriptome in tracheal aspirates. *bioRxiv* (2022).
11. Gomez-Cabrero, D. *et al.* Data integration in the era of omics: Current and future challenges. *BMC Syst. Biol.* **8 Suppl 2**, I1 (2014).
12. Li, B. *et al.* Genome-wide CRISPR screen identifies host dependency factors for influenza a virus infection. *Nat. Commun.* **11**, 164 (2020).
13. Parkinson, N. *et al.* Dynamic data-driven meta-analysis for prioritisation of host genes implicated in COVID-19. *Sci. Rep.* **10**, 22303 (2020).
14. Wang, B. *et al.* Systematic comparison of ranking aggregation methods for gene lists in experimental results. *bioRxiv* (2022).
15. Batra, R. *et al.* Multi-omic comparative analysis of COVID-19 and bacterial sepsis-induced ARDS. *PLoS Pathog.* **18**, e1010819 (2022).
16. Bhargava, M. *et al.* Proteomic profiles in acute respiratory distress syndrome differentiates survivors from non-survivors. *PLoS One* **9**, e109713 (2014).
17. Bhargava, M. *et al.* Bronchoalveolar lavage fluid protein expression in acute respiratory distress syndrome provides insights into pathways activated in subjects with different outcomes. *Sci. Rep.* **7**, 7464 (2017).
18. Bime, C. *et al.* Genome-wide association study in african americans with acute respiratory distress syndrome identifies the selectin P ligand gene as a risk factor. *Am. J. Respir. Crit. Care Med.* **197**, 1421–1432 (2018).

19. Bowler, R. P. *et al.* Proteomic analysis of pulmonary edema fluid and plasma in patients with acute lung injury. *Am. J. Physiol. Lung Cell. Mol. Physiol.* **286**, L1095–104 (2004).

20. Chang, D. W. *et al.* Proteomic and computational analysis of bronchoalveolar proteins during the course of the acute respiratory distress syndrome. *Am. J. Respir. Crit. Care Med.* **178**, 701–709 (2008).

21. Chen, X., Shan, Q., Jiang, L., Zhu, B. & Xi, X. Quantitative proteomic analysis by iTRAQ for identification of candidate biomarkers in plasma from acute respiratory distress syndrome patients. *Biochem. Biophys. Res. Commun.* **441**, 1–6 (2013).

22. Chen, C., Shi, L., Li, Y., Wang, X. & Yang, S. Disease-specific dynamic biomarkers selected by integrating inflammatory mediators with clinical informatics in ARDS patients with severe pneumonia. *Cell Biol. Toxicol.* **32**, 169–184 (2016).

23. Christie, J. D. *et al.* Genome wide association identifies PPFIA1 as a candidate gene for acute lung injury risk following major trauma. *PLoS One* **7**, e28268 (2012).

24. Dolinay, T. *et al.* Inflammasome-regulated cytokines are critical mediators of acute lung injury. *Am. J. Respir. Crit. Care Med.* **185**, 1225–1234 (2012).

25. Dong, H. *et al.* Comparative analysis of the alveolar macrophage proteome in ALI/ARDS patients between the exudative phase and recovery phase. *BMC Immunol.* **14**, 25 (2013).

26. Englert, J. A. *et al.* Whole blood RNA sequencing reveals a unique transcriptomic profile in patients with ARDS following hematopoietic stem cell transplantation. *Respir. Res.* **20**, 15 (2019).

27. Frenzel, J. *et al.* Outcome prediction in pneumonia induced ALI/ARDS by clinical features and peptide patterns of BALF determined by mass spectrometry. *PLoS One* **6**, e25544 (2011).

28. Guillen-Guio, B. *et al.* Sepsis-associated acute respiratory distress syndrome in individuals of european ancestry: A genome-wide association study. *Lancet Respir. Med.* **8**, 258–266 (2020).

29. Howrylak, J. A. *et al.* Discovery of the gene signature for acute lung injury in patients with sepsis. *Physiol. Genomics* **37**, 133–139 (2009).

30. Jiang, Y. *et al.* Single cell RNA sequencing identifies an early monocyte gene signature in acute respiratory distress syndrome. *JCI Insight* **5**, (2020).

31. Juss, J. K. *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.* **194**, 961–973 (2016).

32. Kangelaris, K. N. *et al.* Increased expression of neutrophil-related genes in patients with early sepsis-induced ARDS. *Am. J. Physiol. Lung Cell. Mol. Physiol.* **308**, L1102–13 (2015).

33. Kovach, M. A. *et al.* Microarray analysis identifies IL-1 receptor type 2 as a novel candidate biomarker in patients with acute respiratory distress syndrome. *Respir. Res.* **16**, 29 (2015).

34. Liao, S. Y. *et al.* Identification of early and intermediate biomarkers for ARDS mortality by multi-omic approaches. *Sci. Rep.* **11**, 18874 (2021).

35. Lu, X.-G. *et al.* Circulating miRNAs as biomarkers for severe acute pancreatitis associated with acute lung injury. *World J. Gastroenterol.* **23**, 7440–7449 (2017).

36. Martucci, G. *et al.* Identification of a circulating miRNA signature to stratify acute respiratory distress syndrome patients. *J. Pers. Med.* **11**, 15 (2020).

37. Meyer, N. J. *et al.* ANGPT2 genetic variant is associated with trauma-associated acute lung injury and altered plasma angiopoietin-2 isoform ratio. *Am. J. Respir. Crit. Care Med.* **183**, 1344–1353 (2011).

38. Meyer, N. J. *et al.* IL1RN coding variant is associated with lower risk of acute respiratory distress syndrome and increased plasma IL-1 receptor antagonist. *Am. J. Respir. Crit. Care Med.* **187**, 950–959 (2013).

39. Mirchandani, A. S. *et al.* Hypoxia shapes the immune landscape in lung injury and promotes the persistence of inflammation. *Nat. Immunol.* **23**, 927–939 (2022).

40. Morrell, E. D. *et al.* Cytometry TOF identifies alveolar macrophage subtypes in acute respiratory distress syndrome. *JCI Insight* **3**, (2018).

41. Morrell, E. D. *et al.* Alveolar macrophage transcriptional programs are associated with outcomes in acute respiratory distress syndrome. *Am. J. Respir. Crit. Care Med.* **200**, 732–741 (2019).

42. Nick, J. A. *et al.* Extremes of interferon-stimulated gene expression associate with worse outcomes in the acute respiratory distress syndrome. *PLoS One* **11**, e0162490 (2016).

43. Nguyen, E. V. *et al.* Proteomic profiling of bronchoalveolar lavage fluid in critically ill patients with ventilator-associated pneumonia. *PLoS One* **8**, e58782 (2013).

44. Ren, S. *et al.* Deleted in malignant brain tumors 1 protein is a potential biomarker of acute respiratory distress syndrome induced by pneumonia. *Biochem. Biophys. Res. Commun.* **478**, 1344–1349 (2016).

45. Scheller, N. *et al.* Proviral MicroRNAs detected in extracellular vesicles from bronchoalveolar lavage fluid of patients with influenza virus-induced acute respiratory distress syndrome. *J. Infect. Dis.* **219**, 540–543 (2019).

46. Shortt, K. *et al.* Identification of novel single nucleotide polymorphisms associated with acute respiratory distress syndrome by exome-seq. *PLoS One* **9**, e111953 (2014).

47. Wang, Z., Beach, D., Su, L., Zhai, R. & Christiani, D. C. A genome-wide expression analysis in blood identifies pre-elafin as a biomarker in ARDS. *Am. J. Respir. Cell Mol. Biol.* **38**, 724–732 (2008).

48. Tejera, P. *et al.* Distinct and replicable genetic risk factors for acute respiratory distress syndrome of pulmonary or extrapulmonary origin. *J. Med. Genet.* **49**, 671–680 (2012).

49. Xu, J.-Y. *et al.* Nucleotide polymorphism in ARDS outcome: A whole exome sequencing association study. *Ann. Transl. Med.* **9**, 780 (2021).

50. Zhang, S. *et al.* miR-584 and miR-146 are candidate biomarkers for acute respiratory distress syndrome. *Exp. Ther. Med.* **21**, 445 (2021).

51. Zhang, C. *et al.* Differential expression profile of plasma exosomal microRNAs in acute type a aortic dissection with acute lung injury. *Sci. Rep.* **12**, 11667 (2022).

52. Zhu, Z. *et al.* Whole blood microRNA markers are associated with acute respiratory distress syndrome. *Intensive Care Med. Exp.* **5**, 38 (2017).

53. Christopoulos, D. Introducing unit invariant knee (UIK) as an objective choice for elbow point in multivariate data analysis techniques. *SSRN Electron. J.* (2016).

54. Hu, Y. *et al.* BioLitMine: Advanced mining of biomedical and biological literature about human genes and genes from major model organisms. *G3 (Bethesda)* **10**, 4531–4539 (2020).

55. Quintanilla, E., Diwa, K., Nguyen, A., Vu, L. & Toby, I. T. A data report on the curation and development of a database of genes for acute respiratory distress syndrome. *Front. Genet.* **12**, 750568 (2021).

56. Lachmann, A. *et al.* Massive mining of publicly available RNA-seq data from human and mouse. *Nat. Commun.* **9**, 1366 (2018).

57. Page, M. J. *et al.* The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **372**, n71 (2021).

58. Clark, J. *et al.* A full systematic review was completed in 2 weeks using automation tools: A case study. *J. Clin. Epidemiol.* **121**, 81–90 (2020).

59. Battaglini, D. *et al.* Personalized medicine using omics approaches in acute respiratory distress syndrome to identify biological phenotypes. *Respir. Res.* **23**, 318 (2022).

60. Hernández-Beeftink, T., Guillen-Guio, B., Villar, J. & Flores, C. Genomics and the acute respiratory distress syndrome: Current and future directions. *Int. J. Mol. Sci.* **20**, 4004 (2019).

61. Reilly, J. P., Christie, J. D. & Meyer, N. J. Fifty years of research in ARDS. Genomic contributions and opportunities. *Am. J. Respir. Crit. Care Med.* **196**, 1113–1121 (2017).

62. Zheng, F. *et al.* Novel biomarkers for acute respiratory distress syndrome: Genetics, epigenetics and transcriptomics. *Biomark. Med.* **16**, 217–231 (2022).

63. Christopoulos, D. T. *Inflection: Finds the inflection point of a curve*. (2019).

64. Uhlen, M. *et al.* Towards a knowledge-based human protein atlas. *Nat. Biotechnol.* **28**, 1248–1250 (2010).

65. Uhlen, M. *et al.* A genome-wide transcriptomic analysis of protein-coding genes in human blood cells. *Science* **366**, eaax9198 (2019).

66. Uhlén, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).

67. Dai, Y. *et al.* WebCSEA: web-based cell-type-specific enrichment analysis of genes. *Nucleic Acids Research* **50**, W782–W790 (2022).

68. Raudvere, U. *et al.* G:profiler: A web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* **47**, W191–W198 (2019).

69. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).

70. Gillespie, M. *et al.* The reactome pathway knowledgebase 2022. *Nucleic Acids Res.* **50**, D687–D692 (2022).

71. Martens, M. *et al.* WikiPathways: Connecting communities. *Nucleic Acids Res.* **49**, D613–D621 (2021).

72. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25–29 (2000).

73. Supek, F., njak, M., kunca, N. & muc, T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* **6**, e21800 (2011).

74. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research* **42**, D1001–D1006 (2013).

75. Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research* **44**, W90–W97 (2016).

76. Szklarczyk, D. *et al.* STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).

77. Chin, C.-H. *et al.* cytoHubba: Identifying hub objects and sub-networks from complex interactome. *BMC Systems Biology* **8**, S11 (2014).

78. Shannon, P. *et al.* Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research* **13**, 2498–2504 (2003).

79. Freshour, S. L. *et al.* Integration of the Drug–Gene Interaction Database (DGIdb 4.0) with open crowdsource efforts. *Nucleic Acids Research* **49**, D1144–D1151 (2020).