Dynamic data-driven meta-analysis for prioritisation of host genes implicated in COVID-19

scientific reports

Check for updates

OPEN

Dynamic data-driven meta-analysis for prioritisation of host genes implicated in COVID-19

Nicholas Parkinson[1,4], Natasha Rodgers[1,4], Max Head Fourman[1,4], Bo Wang[1], Marie Zechner[1], Maaike C. Swets[1,2], Jonathan E. Millar[1], Andy Law[1], Clark D. Russell[1,3,5], J. Kenneth Baillie[1,5]✉ & Sara Clohisey[1,5]✉

The increasing body of literature describing the role of host factors in COVID-19 pathogenesis demonstrates the need to combine diverse, multi-omic data to evaluate and substantiate the most robust evidence and inform development of therapies. Here we present a dynamic ranking of host genes implicated in human betacoronavirus infection (SARS-CoV-2, SARS-CoV, MERS-CoV, seasonal coronaviruses). We conducted an extensive systematic review of experiments identifying potential host factors. Gene lists from diverse sources were integrated using Meta-Analysis by Information Content (MAIC). This previously described algorithm uses data-driven gene list weightings to produce a comprehensive ranked list of implicated host genes. From 32 datasets, the top ranked gene was PPIA, encoding cyclophilin A, a druggable target using cyclosporine. Other highly-ranked genes included proposed prognostic factors (*CXCL10*, *CD4*, *CD3E*) and investigational therapeutic targets (*IL1A*) for COVID-19. Gene rankings also inform the interpretation of COVID-19 GWAS results, implicating *FYCO1* over other nearby genes in a disease-associated locus on chromosome 3. Researchers can search and review the gene rankings and the contribution of different experimental methods to gene rank at https://baillielab.net/maic/covid19. As new data are published we will regularly update the list of genes as a resource to inform and prioritise future studies.

There are multiple sources of information that associate host genes with SARS-CoV-2 viral replication, the subsequent host immune response and the ensuing pathophysiology. Integrating these sources of information may provide more robust evidence associating specific genes and proteins with key processes underlying the mechanisms of disease. This is needed in order to make informed judgements about new therapies for inclusion in model studies and clinical trials.

The pace of new research into COVID-19 pathophysiology, including host dependency factors, immune responses, and genetics, has made it nearly impossible to read every report. In addition, assessing the quality and relevance of new evidence is difficult, time-consuming, and requires a high level of expertise. Information from diverse sources has varying quality, scale, and relevance to host responses to SARS-CoV-2. Computational approaches can aid data evaluation and integration. Simple, intuitive methods have a conceptual advantage for translation to decision-making: if both the processes and results are easily comprehensible, then it is easier for human users to trust the conclusions.

SARS-CoV-2 is a betacoronavirus with a 30 kb single-stranded positive-sense RNA genome, and is genetically similar to other human coronaviruses: SARS-CoV, MERS-CoV and the seasonal 'common cold' 229E, OC43, HKU1 and NL63 coronaviruses. Like all viruses, SARS-CoV-2 relies on host machinery to replicate. Host dependency factors represent an attractive target for new therapeutics, as evolution of drug resistance is expected to be slower for host-directed than viral-directed therapies[1].

Treatment directly targeting viral replication can target viral proteins (e.g. remdesivir[2]), or host proteins upon which the virus depends[3]. Host-targeted therapies may have an important role in infectious diseases in general, and the only treatment so far found to reduce mortality in COVID-19—dexamethasone[4]—is likely to act by

[1]Roslin Institute, University of Edinburgh, Easter Bush, Edinburgh EH25 9RG, UK. [2]Department of Infectious Diseases, Leiden University Medical Center, Leiden, The Netherlands. [3]University of Edinburgh Centre for Inflammation Research, The Queen's Medical Research Institute, Edinburgh, UK. [4]These authors contributed equally: Nicholas Parkinson, Natasha Rodgers and Max Head Fourman. [5]These authors jointly supervised this work: Clark D. Russell, J.Kenneth Baillie and Sara Clohisey. ✉email: j.k.baillie@ed.ac.uk; sara.clohisey@roslin.ed.ac.uk

targeting host immune-mediated organ damage[5]. Other host-directed treatments (e.g. anakinra, tocilizumab, sarilumab, mavrilimumab), repurposed from other indications, are currently under investigation[2,6–10].

In this analysis we systematically identify and combine existing data from human betacoronavirus research to generate a comprehensive ranked list of host genes involved in COVID-19. We comment on the application of this resource to inform further research on COVID-19 pathogenesis and prioritise host therapeutic targets.

To identify existing literature which could provide informative datasets for host gene prioritisation, we conducted a systematic review of published studies and pre-print manuscripts pertaining to host gene involvement in human betacoronavirus infection and associated disease. Results from identified studies, in the form of lists of implicated host factor genes, were combined using meta-analysis by information content (MAIC)[3], an approach we previously developed to identify host genes necessary for Influenza A virus (IAV) replication. We have previously demonstrated that the MAIC algorithm successfully predicts new experimental results from an unseen future experiment[3]. Our gene prioritisation results both recapitulate existing understanding of COVID-19 pathophysiology and highlight key host factors and potential therapeutic targets that have to date been largely overlooked.

Methods

Meta-analysis by information content (MAIC). MAIC allows the combination of data from diverse sources without prior assumptions regarding the quality of each individual data source. The MAIC approach begins with the following assumptions:

- There exists a set of true positives: host genes involved in COVID-19 pathogenesis.
- A gene is more likely to be a true positive if it is found in multiple experiments.
- A gene is more likely to be a true positive if it occurs in a list containing a higher proportion of genes with supporting evidence from multiple sources.
- Due to experimental biases, the evidence that a gene is a true positive is further increased if it is found across experimental types.

With these assumptions, MAIC allows the quantification of the information content in a gene list by comparing that list to the results from other experiments that might reasonably be expected to find some of the same genes. Input gene lists can be categorised by data type (Table 2), allowing comparison both within and between methodologies. MAIC produces a weighting factor for each experiment, and this weighting is used to calculate a score for each gene (Fig. 1A). The analysis then produces a final ranked list of genes based on this score, which summarises the combined evidence from all input sources of that particular gene being involved in SARS-CoV-2 pathogen-host interaction. A full description of the MAIC algorithm can be found in our original report[3].

Data eligibility. Inclusion and exclusion criteria are shown in Table 1. To complement emerging data pertaining to the novel SARS-CoV-2, we included studies of other human coronaviruses. Included methodologies are shown in Table 2. We reduced the bias that can be caused by focussing on specific genes, by excluding candidate gene or single gene studies.

Literature search. A systematic literature search of PubMed was conducted on 28/04/2020 and updated weekly until 06/07/2020. We used the following search strategy, with no date or language restrictions: > Keywords: ("Coronavirus" OR "Severe Acute Respiratory Syndrome" OR "Middle East Respiratory Syndrome" OR "Sars-CoV-2" OR "COVID-19") AND ((gene*.[Title/Abstract]) OR (genome*[Title/Abstract]) OR (transcript*[Title/Abstract]) OR (protein*[Title/Abstract]) OR ("Susceptibility"[Title/Abstract]) OR (siRNA[All Fields])).

Potentially relevant pre-print manuscripts were identified by screening all papers categorised as COVID-19-related in the bioRxiv and medRxiv servers. Titles and abstracts of all returned papers were first assessed for relevance and duplication by a single member of the review team. Following this, full-length texts were obtained and an in-depth review was carried out by two further reviewers, independently, in order to confirm eligibility according to Tables 1 and 2. In cases where a consensus was not reached, a third reviewer appraised the paper. This method ensured each paper was assessed for eligibility by a minimum of three independent reviewers. Relevant data, as shown in Table 3, was extracted from each reviewed paper.

Gene list extraction and categorisation. Relevant gene lists were identified and extracted. Datasets were excluded from the analysis where insufficient data were available to construct a meaningful unbiased gene list, for example where results for only a non-systematically selected subset of genes of interest were reported. Gene lists were categorised based on methodology as shown in Table 2.

Gene list rankings were preserved where possible, if sufficient numerical data were available. Rankings were based on significance or magnitude of effect. Adjusted measures of significance, usually adjusted p, were prioritised over raw p and $logFC$ to determine ranking where multiple values were available. For studies reporting comparisons at multiple time points, genes were ranked based on the minimum p across all comparisons. To exclude irrelevant genes, a significance or effect size threshold was applied to all lists. This was either the threshold used by the authors for reporting, or where full data were provided this was determined as adjusted $p < 0.05$, $|z\ score| > 1.96$ or $|logFC| > 1.5$ depending on available values.

Gene, transcript and protein names or identification numbers were converted to the associated HGNC gene symbol, or an equivalent Ensembl or Refseq symbol where no HGNC symbol existed. Non-primate genes were

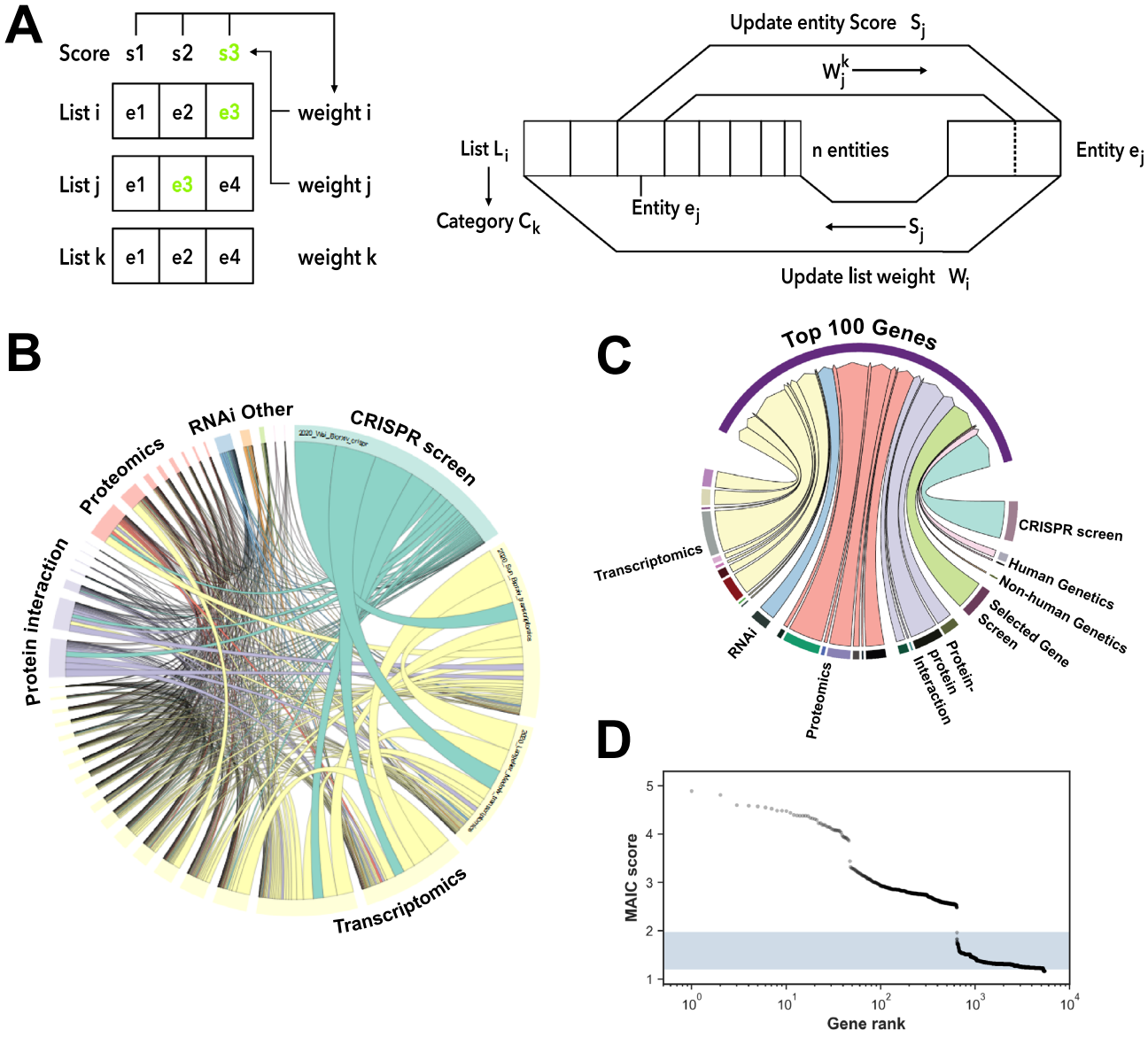

Figure 1. Overview of MAIC approach. (**A**) Schematic showing the operation of MAIC. Each entity in a list is given a score, based on overlap with other lists and rank where relevant, and each list is given a weight determined by the scores of its constituent entities. Entity scores are iteratively updated using list weights, and list weights are updated using entity scores, until convergence occurs. (**B**) Circular plot showing overlap between different data sources included in MAIC. Size of data source blocks is proportional to the summed information content (MAIC scores) of the input list. Lines are coloured according to the dominant data source. Data source categories share the same colour; the largest categories and data sources are labelled (see Supplementary Information for full source data). (**C**) Relative information contributions (determined by sum of MAIC score contributions) of each experimental category to the evidence base for the top 100 genes in the MAIC output. (**D**) Distribution of MAIC scores by gene rank. The shaded region indicates the range of possible scores for a gene supported by a single gene list only. Beyond ranks around 700 in this study, gene scores approach baseline, indicating they have little corroborative evidence.

mapped to their human homologues using the NCBI Homologene database[11], or excluded from the analysis if no human homologue could be identified.

Gene set enrichment analysis. Rank-based gene set enrichment analysis was performed using the package FGSEA in R version 3.5.2, with genes ranked by MAIC score[12]. *p*-values were estimated using an empirical probability distribution based on 10^6 permutations. Two additional methods were used for comparison. Flexible-threshold analysis of minimum hypergeometric scores was conducted using the full ranked list, using GOrilla (for Gene Ontology terms only)[13]. Gene set over-representation in the top 100 genes was analysed by a Fisher's exact test as implemented in Enrichr[14]. The Benjamini–Hochberg procedure was used to control the false discovery rate ($FDR < 0.05$) for all methods.

Inclusion	Exclusion
Infection of any species with SARS-CoV, SARS-CoV-2, MERS-CoV, HCoV-229E, HCoV-OC43, HCoV-HKU1 or HCoV-NL63	Candidate in vitro or in vivo gene, transcript or protein studies and screens—defined here as < 50 genes, transcripts or proteins investigated
Human studies: in vivo, in vitro, primary human cells, in vitro human cell lines	Candidate-gene human genetic studies
Animal studies: in vivo, ex vivo, in vitro, primary cells, in vitro cell lines	< 5 hosts in virus group or control group in patient studies
Accepted experimental designs in Table 2	Meta-analyses, in silico anayses, re-analysis of data published elsewhere
	Insufficient data available

Table 1. Entry criteria.

Accepted methodologies	MAIC category
CRISPR screen	CRISPR Screen
RNAi screen	RNAi
Protein–protein interaction e.g. yeast-2-hybrid screen	Protein–protein interaction
Host proteins incorporated into virion or virus like particle	Virus
Genetic Association Studies Human	Human genetics
Genetic Association Studies Non-human	Non-human genetics
Proteomic studies e.g. mass-spectrometry	Proteomics
Selected gene set screens	Gene set screen

Table 2. Methodologies accepted for inclusion in meta-analysis and associated labels.

Extracted information	Examples
Virus & virus component/modification	SARS-CoV-2, HCoV-229E
Method/experimental design	See Table 2
Organism	Human, rodent, Non-human primate
Cell/tissue type	Vero6, A549, serum
Peer reviewed or pre-print	Peer-reviewed, pre-print

Table 3. Data extracted from each publication.

Results

Systematic review of the literature. We identified a total of 31 studies with available data meeting our eligibility criteria (12 pre-print manuscripts and 19 peer-reviewed studies), yielding 32 gene lists (Supplementary Fig. S1, Supplementary Table 1). The included gene lists comprised 11 ranked and 21 unranked lists, in 8 experimental categories, with list lengths ranging from three to 9,967 genes (median 61). The datasets included 3 genetic perturbation screens (CRISPR, RNAi and interferon-stimulated gene overexpression), 3 genetic studies (of which 2 were in humans), 7 protein–protein interaction studies, 7 proteomic and 12 transcriptomic studies.

MAIC analysis of identified studies. Of 5418 genes implicated in human betacoronavirus infection in these 32 datasets, 4150 are supported by a single paper only, 629 had evidence from more than one source within the same experimental category, and 639 are supported by data from multiple study types. Although extensive within-category overlap was seen for transcriptomic studies, there was less concordance within categories such as proteomics and protein–protein interaction. As with our previous study of influenza, contributions from one CRISPR screen dominate the overall information content (Fig. 1B). This was due in part to list length and hence contribution to scores for multiple lower-ranking genes. Information contributions for the top 100 genes are more balanced (Fig. 1C). The MAIC score distribution (Fig. 1D) reflects the degree of cross-category overlap, with the highest ranked genes supported by data from three distinct experimental categories.

The highest ranking genes and their contributing evidence sources are shown in Fig. 2A and Supplementary Table 2. Top genes include *IL1A*[15] and other components of the innate and adaptive immune systems (such as *CXCL10*, *CD4* and *TLR3*), which have previously been shown to contribute to COVID-19 pathogenesis. Other top genes have not previously received much attention in the context of coronavirus infection. These include *PPIA* (cyclophilin A), and *RYBP* (RING1 and YY1 binding protein), which play roles in protein folding, transcriptional repression, regulation of proteasomal degradation and apoptosis.

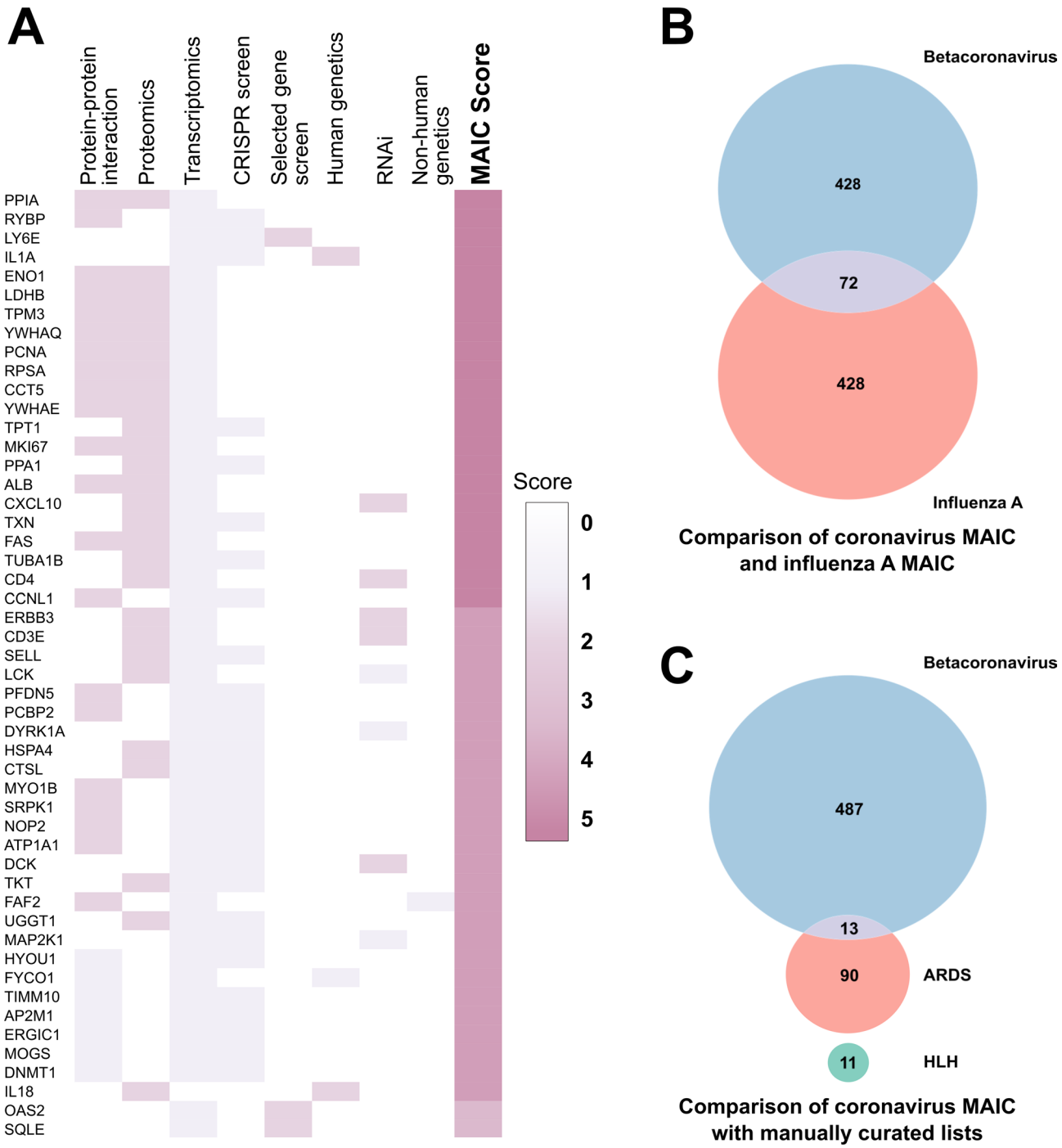

Figure 2. Highest ranked genes in the MAIC output and overlap with other conditions. (**A**) Heatmap of the top 50 genes implicated in SARS-CoV-2 infection, as ranked by the MAIC algorithm. The heatmap shows the information sources contributing to each of the top genes, by experimental category. Full details of all scored genes, including specific studies contributing to each, are given in Supplementary Table S1. (**B**) Venn diagram of overlap between the top 500 hits from this study and the top 500 hits from our previous MAIC analysis of Influenza A virus. (**C**) Venn diagram of overlap between the top 500 hits from this study and manually curated lists from available literature on HLH and ARDS.

An up-to-date prioritised list of implicated genes is available at baillielab.net/maic/covid19. We will repeat the analysis regularly as new data become available.

Overlap of MAIC output with respiratory disease and HLH associated genes. Death in severe COVID-19 is usually a consequence of lung injury leading to ARDS (Acute Respiratory Distress Syndrome), a final common pathway that can occur in any severe acute respiratory infection. Host susceptibility factors in

COVID-19 may be shared with other infections or ARDS. We compared the output from our analysis to our previous MAIC analysis of Influenza A virus[3]. Among the top 500 ranked genes from each output, we found 72 overlapping genes (Fig. 2B), including a large number of RNA-binding, ribosome-associated and chaperone genes. Unexpectedly few immune related genes overlapped. This is surprising as both viruses are single-stranded RNA viruses and despite differing in sense, intracellular pathogen detection mechanisms are expected to be similar.

To expand this analysis we manually curated genes associated with ARDS from previously published literature reviews and determined the overlap with our MAIC output (Supplementary Table 4). Among the top 500 ranked genes from the coronavirus MAIC output we found an overlap of 13 genes with the ARDS list (consisting of 103 genes) (Fig. 2C). Here we saw a number of genes associated with innate immunity and modulation of inflammation (*TNFα*, *IL6*, *IL18*, *CCL2*, *IL1B*, *TLR1*, *IL13*, *NFκBIA*). This relatively small overlap was also observed in a similar analysis comparing MAIC output to a gene list curated as part of a published review looking at genes implicated in respiratory disease as a consequence of infection (Supplementary Fig. S2A)[16].

The inflammatory profile, including hyperferritinaemia, observed in COVID-19 has led to the suggestion that a form of secondary haemophagocytic lymphohistiocytosis (HLH), a hyper-inflammatory syndrome, could be occurring[17]. We manually curated genes involved in the familial form of this syndrome, based on previously published review articles and mutations that are tested for in clinical practice (Supplementary Table 4), and compared these with our MAIC output, finding no overlap, although only eleven genes were found in the literature to be associated with familial HLH.

Finally, we also performed a comparison of our output with a recently published systematic review also identifying genes implicated in betacoronavirus infection and focused on peer-reviewed articles concerned with biomarkers associated with a clinical diagnosis of SARS or associated syndromes[18]. This review identified 22 unique genes, 6 of which overlap with the MAIC output and are detailed in Supplementary Fig. S2B.

Pathway analysis of MAIC output. To better understand the biological functions of the most strongly implicated genes, we performed gene set enrichment analysis in ten databases of functional annotations. We used three complementary methods, assessing enrichment either in terms of rank distribution across the whole dataset (permissive) or in over-representation in the top 100 genes only (conservative). There was extensive overlap between these approaches (Fig. 3A and Supplementary Table 3).

Functional annotations that were significant using at least two methods and that were reflected in results from more than one database included terms related to cytokine, toll-like receptor and T-cell receptor signalling, protein processing, apoptosis, the complement system, VEGF signalling, glucose metabolism and viral infections such as influenza A. As expected, the relative information contributions from different experiment types varied between pathways (Fig. 3B). For example, terms related to complement and coagulation received relatively little contribution from CRISPR screen data (derived from epithelial cells) but more information from proteomics and experiments using serum or swab samples, whilst pathways related to protein processing had relatively greater contributions from protein interaction studies. In all cases, enriched pathways drew information from a range of experiment types.

Integration of MAIC output with results from published GWAS. One of the principal applications of MAIC is in the interpretation of results of genome-wide association studies (GWAS). Genome-wide association studies often implicate a locus containing a number of candidate genes and the precise nature of the interaction between gene and disease may not be known. As an example, we applied our results to the locus in chromosome 3 associated with hospitalisation due to COVID-19 in the sole COVID-19 GWAS published at the time of writing (data from which were also included in this analysis)[19]. This locus contained six genes (*SL6A20*, *LTZFL1*, *CCR9*, *FYCO1*, *CXCR6* and *CXR1*) that could all plausibly be linked to COVID-19 pathophysiology on the basis of their known functions.

Of these, *FYCO1*, which encodes a protein involved in vesicle transport and autophagy, was highly ranked in our results (rank 42). *FYCO1* is supported by SARS-CoV-2-specific protein–protein interaction and transcriptomic data. *CCR9* (rank 417) had additional support from a single transcriptomic study, while *SL6A20*, *LTZFL1*, *CXCR6* and *XCR1* had low ranks in our results, with no corroborating evidence in other studies.

Discussion

The interpretation of any meta-analysis is critically dependent on the criteria for inclusion. In this case, our objective is to cast the net wide, including a range of data sources that are both conceptually and methodologically divergent. Experimental results bearing little relation to the composite of evidence from the other studies are downgraded by the MAIC algorithm, so the effect of irrelevant, noisy, or poorly-conducted experiments is minimal[3]. By using permissive inclusion criteria, together with weighted meta-analysis, we have identifed key elements of the host–pathogen interaction and promising therapeutic targets for further investigation and intervention. These include host factors involved in viral replication, and elements of the immune response, which have been overlooked in the contributing studies (Fig. 4). We contend that the output of our analysis can be applied to (1) inform understanding of pathogenesis and planning of in vitro and in vivo validation studies of selected host factors; (2) prioritise host therapeutic targets, through matching highly ranked host factors with available drugs; and (3) inform the interpretation of hits from GWAS (for example in prioritising candidates for further investigation within the chromosome 3 locus discussed above) and studies of monogenic inborn errors of immunity.

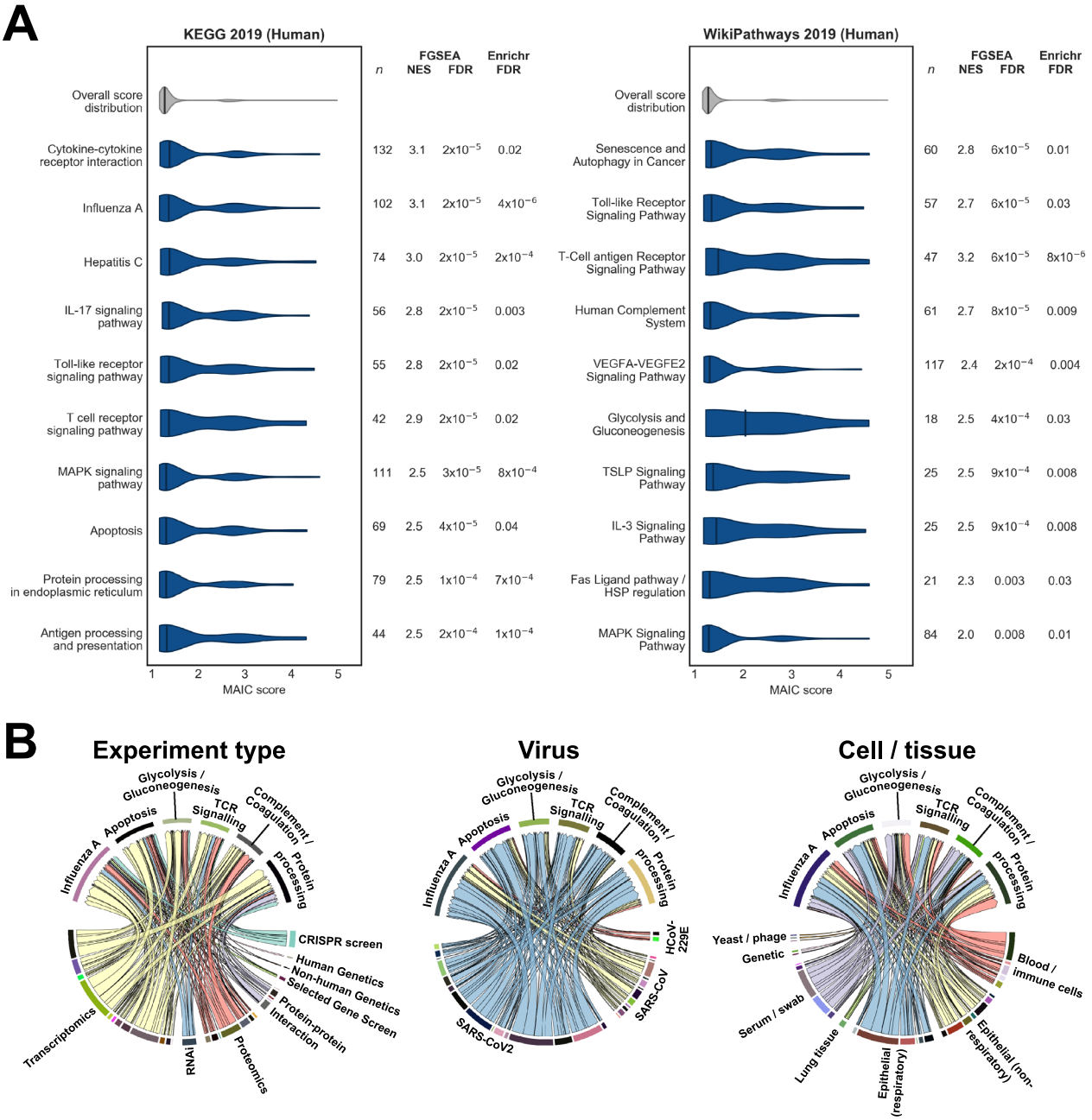

Figure 3. Gene Set Enrichment Analysis of MAIC rankings. (**A**) Violin plots of MAIC score distributions of top enriched pathways significant with both FGSEA and Enrichr algorithms, from the KEGG 2019 (Human) and WikiPathways 2019 (Human) databases. Highly similar pathways and irrelevant specific disease terms are not shown. n: number of gene set members included in the overall MAIC output; NES: normalised enrichment score from FGSEA. (**B**): Information contribution by methodology for selected enriched KEGG terms. Relative contributions of different information sources vary between functional annotations, but no single methodology predominates to drive enrichment.

Key host factors related to viral entry and replication. Coronaviruses hijack host endomembranes to facilitate anchoring of the replication/transcription complex[20]. Consistent with this, we observe an over-representation of endoplasmic reticulum-related genes. Genes related to the function of the endoplasmic reticulum (*FAF2*, *ERGIC1*, *TENT5C*, *TENT5A*, *CFL1*, *STARD5*) and glycosylation (*MOGS*, *UGDH*, *UGGT1*, *PDIA6*, *PDIA3*) are observed along with a number of chaperone proteins (*HSPB1*, *HSPA4*, *HSPA8*, *HSP90AA1*, *ST13*). Many of these genes are related to the unfolded protein response (UPR), a stress response initiated by accumulation of misfolded proteins. *FYCO1*, implicated in published GWAS results, has been suggested as a key mediator linking ER-derived double membrane vesicles, the primary replication site for coronaviruses, with the microtubule network[21].

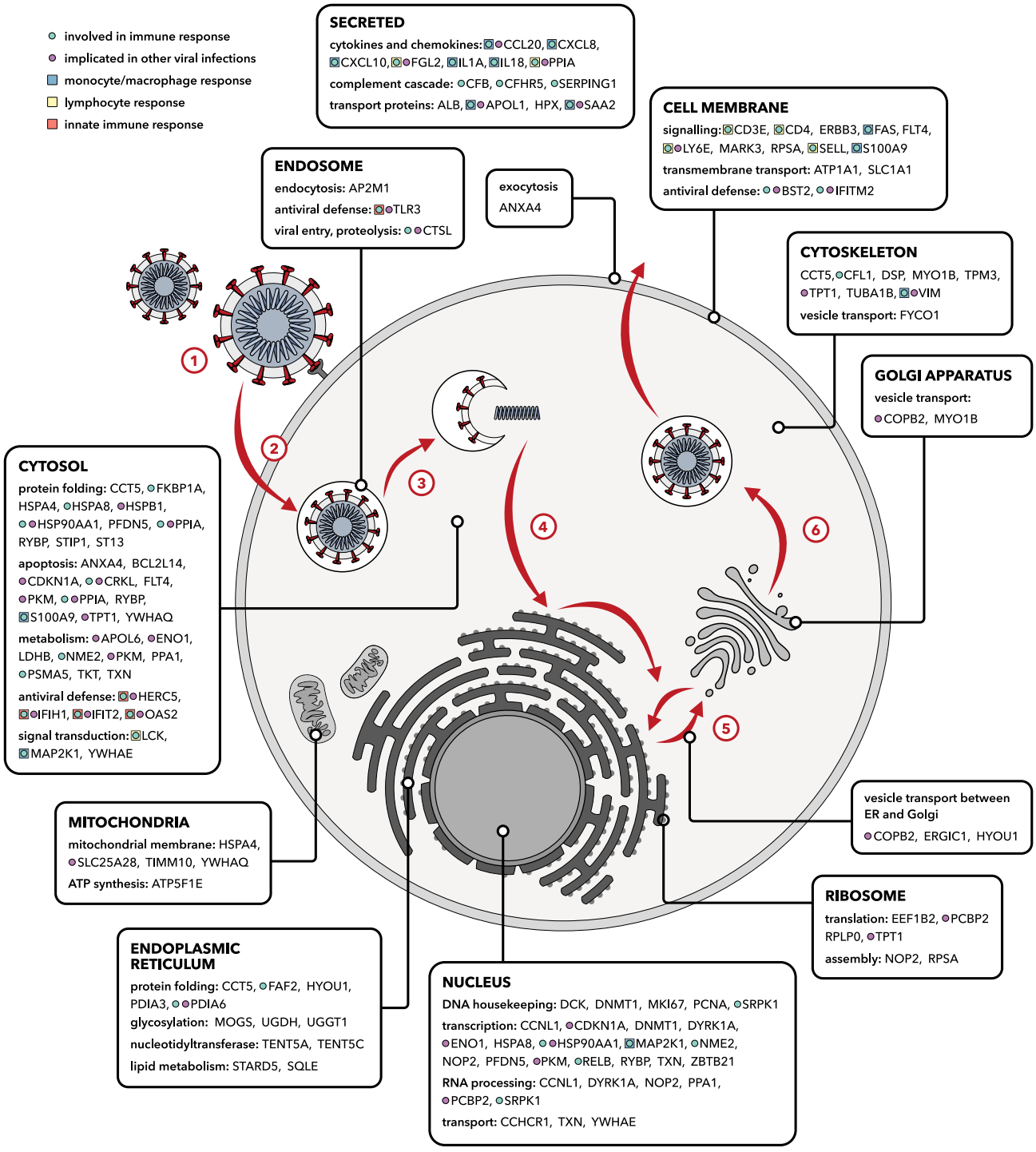

Figure 4. Cellular functions of the 100 highest ranked genes in the MAIC output. Protein products of these genes have diverse cellular locations and are associated with numerous processes relevant to the viral life cycle and host immune system. Stages of the betacoronavirus life cycle: (1) S protein-mediated attachment to the cell surface. (2) Endocytosis. (3) Membrane fusion and viral genome release into the cytoplasm. (4) Assembly of the replication-transcription complex, translation of mRNA. (5) Viral replication and virion assembly. (6) Virion maturation, budding and translocation of vesicles.

Viral entry via spike (S) protein-mediated membrane fusion is well characterised[22]. The first step, spike activation, requires cleavage via host proteases such as cathepsin L (*CTSL*, rank 31)[23]. In the Wei et al.[24] CRISPR screen, knockout of *CTSL* restricted viral production. Cathepsin L inhibition has been suggested as a promising therapeutic strategy for COVID-19: specific small molecule inhibitors are in early stages of development, and direct or indirect inhibition is observed with a number of approved drugs including glycopeptide antibiotics, chloroquine and dexamethasone[25]. *ATP1A1* (rank 35), encoding a subunit of the NA + /K + cotransporter, has

similarly been shown to be necessary for membrane fusion and viral entry for a number of coronaviruses[26]. Inhibition of *ATP1A1* by cardiac glycosides suppressed MERS-CoV infection in vitro[27]. Additional anti-inflammatory[28] effects of these drugs make them theoretically attractive therapeutic options, but adverse effects may limit their practical application.

The interferon-stimulated gene *LY6E* (rank 3), which plays a key role in enhancing cellular entry by RNA viruses including influenza A virus[29], was unexpectedly found to have a strong restricting effect on SARS-CoV-2, SARS-CoV and MERS-CoV[30]. Such widely opposite differential effects on different viruses have been reported for other host genes, such as *IFITM3*, *RSAD2* and *AXL* (ranked 1056, 1039 and 215 respectively by MAIC)[29], and for *PPIA* (see below).

Immune response to SARS-CoV-2. Consistent with the emerging understanding of the pathogenesis of COVID-19, key genes in the inflammatory response to SARS-CoV-2 infection are highly represented in the top 100 genes. These include genes involved in recognizing the virus (*TLR3*, *IFIH1*), activating the innate immune system (*OAS2*, *HERC5*, *S100A9*), chemotaxis (*S100A9*, *CXCL10*, *CXCL8*, *CCL20*, *SAA2*) and pro-inflammatory cytokines (*IL1A*, *IL18*). Toll-like receptor 3 (*TLR3*) is an endosome-associated pathogen-associated molecular pattern receptor, constitutively expressed in the respiratory tract and many immune cells. TLR3 detects double-stranded viral RNA and triggers production of type I interferons and other pro-inflammatory cytokines, such as IL6 (rank 104) and TNFα (rank 182) via IRF3 and NF-κB[31].

The chemokine CXCL10 (rank 17) is a key signalling molecule in viral immunity, which could contribute to pulmonary inflammation as well as aiding viral clearance[32]. CXCL10 levels are associated with outcome in influenza[33] and are thought to have a protective effect in SARS, but a pro viral effect in HIV[34]. CXCL10 has been proposed as a prognostic marker for the progression of disease in COVID-19, with continuously high levels of CXCL10 associated with worse outcomes[35].

The high rankings of genes associated with activation and binding of T lymphocytes (*CD4*, *CD3E*, *FGL2*, *LCK*, *SELL*) are also likely related to their prognostic significance, as lymphopaenia is strongly associated with poor outcomes in COVID-19[36,37]. Absolute counts of CD3+, CD4+ and CD8 + T lymphocytes have been proposed as a potential predictor of outcome in severe COVID-19 patients[38] with an increase in numbers of these cells observed during recovery.

Prioritisation of host susceptibility factors as therapeutic targets. The highest-ranking gene is *PPIA*, which encodes peptidyl-prolyl cis–trans isomerase A (*PPIA*, also known as cyclophilin A, *CypA*), a cytosolic protein involved in protein folding and trafficking, cell signalling and T-cell activation via the calcineurin/NFAT pathway[39]. PPIA is a pro-viral factor for hepatitis C virus (HCV), HIV-1, and SARS-CoV, and an anti-viral factor for IAV[40,41].

The cyclophilin inhibitor cyclosporine has in vitro antiviral activity against HCV[42–44]. This was also observed in a HCV clinical trial, where cyclosporine combined with interferon-α was more efficacious in achieving sustained virologic response than interferon monotherapy[45]. Similar in vitro and clinical results were demonstrated for the PPIA inhibitor alisporivir (DEBIO-015)[46]. PPIA is also a pro-viral factor for HIV-1 and alisporivir can inhibit HIV-1 replication in vitro[47,48].

A genome-wide protein–protein interaction screen identified an interaction between the SARS-CoV Nsp1 protein and PPIA[49]. Cyclosporine inhibits SARS-CoV replication in Vero E6 cells, as well as HCoV-229E, HCoV-NL63, avian coronavirus and feline coronavirus. Nsp1 also induced IL-2 expression in HEK293 cells through the calcineurin/NFAT pathway, making inhibition of this pathway interesting from both an antiviral and immunomodulatory perspective[50].

Two interleukin 1 superfamily members, *IL1A* and *IL18*, were in the top 50 ranked genes. The high ranking of *IL1A* (encoding interleukin 1-α) is striking because monoclonal antibodies against interleukin 1 receptor are a plausible therapeutic approach for COVID-19[9]. This pro-inflammatory cytokine, which is synergistic with TNFα, is constitutively expressed in epithelial cells and is upregulated after SARS-CoV-2 infection[51]. Interleukin-1 receptor blockade with anakinra is now being tested in a number of randomised clinical trials in COVID-19[9]. The pro-inflammatory cytokine IL18 (interleukin 18) is also of potential therapeutic relevance. IL18 is a product of the activated NLRP3 inflammasome and circulating levels are elevated in Covid-19 and positively correlate with severity[52]. Therapeutic inhibition has been investigated in patients with adult-onset Still's disease using recombinant human IL-18 binding protein (tadekinig alfa) which appears safe and potentially efficacious[53]. Small molecule inhibitors are also in development[54].

Advantages and limitations of data integration via MAIC. The principal advantage of the MAIC approach is that it allows integration of data from diverse sources. Unlike other methods for gene list comparison such as vote counting or robust rank aggregation[55], MAIC applies a data-driven weighting to each dataset, accepts both ranked and unranked lists, and includes user-defined categories which prevent any single method from overwhelming the results. MAIC outperforms other methods for predicting antiviral genes[3].

This meta-analysis is restricted to studies involving genome-wide hypotheses or screening data for large gene sets, and does not consider evidence from candidate gene genetic studies or single-gene perturbations. Where a single gene has been investigated extensively but genome-scale studies are sparse, our approach may underestimate the relative strength of evidence for certain genes. Single gene studies, however, are likely to focus preferentially on genes that fit pre-conceived ideas of disease pathogenesis and may be prone to other biases such as publication bias, something which we mitigated against in our inclusion criteria.

Genetic perturbation data are still relatively sparse for SARS-CoV-2 and other human betacoronaviruses: only one genome-wide CRISPR knockout screen and two other sub-genome-scale screens (kinome-wide RNAi

and interferon-stimulated gene overexpression screens) were included in the meta-analysis. Limited data of this type could be responsible for the lower than expected rankings for *ACE2* (rank 320), a major functional receptor for the SARS-CoV and SARS-CoV-2 spike (S) proteins, and *TMPRSS2* (rank 3037), a serine protease required for S protein priming[22,56,57]. While *ACE2* was identified as a host dependency factor in the CRISPR screen, *TMPRSS2* was not, and as neither gene was included in the other two screens, the effects (or lack thereof) could not be confirmed. The only other supportive evidence for a role in disease pathophysiology, in studies included here, came from a single transcriptomic study for each; there was no evidence from protein–protein interaction, proteomics or genetics[51,58]. Candidate gene association studies are highly dependent on prevalence of functional variants and, while overlapping to a limited extent with results of large-scale screens included here, have been notably unable to detect significant associations with *ACE2*[18]. Integrating perturbation data will thus add considerably to our ability to interpret the relative importance of these factors. Both *ACE2* and *TMPRSS2* have been proposed as possible therapeutic targets for COVID-19, and clinical trials are underway for the TMPRSS2 inhibitors nafamostat and camostat mesylate.

Systematic review and meta-analysis are routine elements in the assessment of clinical evidence and some fields in genomics, but have been less widely applied to mechanistic biology. Using a flexible and intuitive method, we have systematically reviewed and meta-analysed host gene-level data from studies that address a range of complementary questions regarding human betacoronavirus infection. This provides external validation for numerous host genes implicated in both in the viral life cycleand in the immune response and identifies several plausible therapeutic targets with broad support from multiple sources. As more, and larger, datasets become available we expect the accuracy of MAIC will improve with each iteration.

Received: 24 September 2020; Accepted: 2 December 2020
Published online: 18 December 2020

References

1. Baillie, J. K. Translational genomics. Targeting the host immune response to fight infection. *Science* **344**, 807–808 (2014).
2. Beigel, J. H. *et al.* Remdesivir for the treatment of covid-19: preliminary report. *N. Engl. J. Med.* https://doi.org/10.1056/NEJMoa2007764 (2020).
3. Li, B. *et al.* Genome-wide CRISPR screen identifies host dependency factors for influenza A virus infection. *Nat. Commun.* **11**, 164 (2020).
4. Horby, P. *et al.* Dexamethasone in hospitalized patients with covid-19 - preliminary report. *N. Engl. J. Med.* https://doi.org/10.1056/NEJMoa2021436 (2020).
5. Dorward, D. A. *et al.* Tissue-specific immunopathology in fatal Covid-19. *Am. J. Respir. Crit. Care Med.* https://doi.org/10.1164/rccm.202008-3265OC (2020).
6. Luca, G. D. *et al.* GM-CSF blockade with mavrilimumab in severe COVID-19 pneumonia and systemic hyperinflammation: a single-centre, prospective cohort study. *Lancet Rheumatol.* **2**, e465–e473 (2020).
7. Wang, Y. *et al.* Remdesivir in adults with severe covid-19: a randomised, double-blind, placebo-controlled, multicentre trial. *Lancet* **395**, 1569–1578 (2020).
8. Cao, B. *et al.* A trial of lopinavir-ritonavir in adults hospitalized with severe covid-19. *N. Engl. J. Med.* **382**, 1787–1799 (2020).
9. Cavalli, G. *et al.* Interleukin-1 blockade with high-dose anakinra in patients with covid-19, acute respiratory distress syndrome, and hyperinflammation: a retrospective cohort study. *Lancet Rheumatol.* **2**, e325–e331 (2020).
10. Guaraldi, G. *et al.* Tocilizumab in patients with severe COVID-19: a retrospective cohort study. *Lancet Rheumatol.* **2**, e474–e484 (2020).
11. Agarwala, R. *et al.* Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **44**, D7-19 (2016).
12. Sergushichev, A. A. An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. *bioRxiv* https://doi.org/10.1101/060012 (2016).
13. Eden, E., Navon, R., Steinfeld, I., Lipson, D. & Yakhini, Z. GOrilla: a tool for discovery and visualization of enriched go terms in ranked gene lists. *BMC Bioinform.* **10**, 48 (2009).
14. Chen, E. Y. *et al.* Enrichr: Interactive and collaborative html5 gene list enrichment analysis tool. *BMC Bioinform.* **14**, 128 (2013).
15. Ong, E. Z. *et al.* A dynamic immune response shapes covid-19 progression. *Cell Host Microbe* **27**, 879–882 (2020).
16. Patarčić, I. *et al.* The role of host genetic factors in respiratory tract infectious diseases: systematic review, meta-analyses and field synopsis. *Sci. Rep.* **5**, 16119 (2015).
17. Mehta, P. *et al.* COVID-19: consider cytokine storm syndromes and immunosuppression. *Lancet* **395**, 1033–1034 (2020).
18. Di Maria, E., Latini, A., Borgiani, P. & Novelli, G. Genetic variants of the human host influencing the coronavirus-associated phenotypes (sars, mers and covid-19): Rapid systematic review and field synopsis. *Hum. Genom.* **14**, 30 (2020).
19. Ellinghaus, D. *et al.* Genomewide association study of severe covid-19 with respiratory failure. *N. Engl. J. Med.* https://doi.org/10.1056/NEJMoa2020283 (2020).
20. Knoops, K. *et al.* SARS-coronavirus replication is supported by a reticulovesicular network of modified endoplasmic reticulum. *PLoS Biol.* **6**, e226 (2008).
21. Reggiori, F., de Haan, C. A. M. & Molinari, M. Unconventional use of lc3 by coronaviruses through the alleged subversion of the erad tuning pathway. *Viruses* **3**, 1610–1623 (2011).
22. Li, W. *et al.* Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus. *Nature* **426**, 450–454 (2003).
23. Bosch, B. J., Bartelink, W. & Rottier, P. J. M. Cathepsin l functionally cleaves the severe acute respiratory syndrome coronavirus class i fusion protein upstream of rather than adjacent to the fusion peptide. *J. Virol.* **82**, 8887–8890 (2008).
24. Wei, J. *et al.* Genome-wide CRISPR screen reveals host genes that regulate SARS-CoV-2 infection. *J. Virol.* https://doi.org/10.1101/2020.06.16.155101 (2020).
25. Liu, T., Luo, S., Libby, P. & Shi, G.-P. Cathepsin l-selective inhibitors: a potentially promising treatment for covid-19 patients. *Pharmacol. Ther.* https://doi.org/10.1016/j.pharmthera.2020.107587 (2020).
26. Burkard, C. *et al.* ATP1A1-mediated src signaling inhibits coronavirus entry into host cells. *J. Virol.* **89**, 4434–4448 (2015).
27. Ko, M. *et al.* Screening of FDA-approved drugs using a MERS-CoV clinical isolate from South Korea identifies potential therapeutic options for COVID-19. *bioRxiv* https://doi.org/10.1101/2020.02.25.965582 (2020).
28. Ihenetu, K. *et al.* Digoxin and digoxin-like immunoreactive factors (dlif) modulate the release of pro-inflammatory cytokines. *Inflamm. Res.* **57**, 519–523 (2008).
29. Mar, K. B. *et al.* LY6E mediates an evolutionarily conserved enhancement of virus infection by targeting a late entry step. *Nat. Commun.* **9**, 3603 (2018).

30. Pfaender, S. *et al.* LY6E impairs coronavirus fusion and confers immune control of viral disease. *bioRxiv* https://doi.org/10.1101/2020.03.05.979260 (2020).
31. Vercammen, E., Staal, J. & Beyaert, R. Sensing of viral infection and activation of innate immunity by toll-like receptor 3. *Clin. Microbiol. Rev.* **21**, 13–25 (2008).
32. Birra, D. *et al.* COVID 19: a clue from innate immunity. *Immunol. Res.* **68**, 161–168 (2020).
33. Dunning, J. *et al.* Progression of whole-blood transcriptional signatures from interferon-induced to neutrophil-associated patterns in severe influenza. *Nat. Immunol.* **19**, 625–635 (2018).
34. Liu, M. *et al.* CXCL10/ip-10 in infectious diseases pathogenesis and potential therapeutic implications. *Cytokine Growth Factor Rev.* **22**, 121–130 (2011).
35. Coperchini, F., Chiovato, L., Croce, L., Magri, F. & Rotondi, M. The cytokine storm in covid-19: an overview of the involvement of the chemokine/chemokine-receptor system. *Cytokine Growth Factor Rev.* **53**, 25–32 (2020).
36. Huang, I. & Pranata, R. Lymphopenia in severe coronavirus disease-2019 (covid-19): systematic review and meta-analysis. *J. Intens. Care* **8**, 36 (2020).
37. Tan, L. *et al.* Lymphopenia predicts disease severity of covid-19: a descriptive and predictive study. *Signal Transd. Target. Therapy* **5**, 33 (2020).
38. Wang, J. *et al.* Characteristics of lymphocyte subsets and their predicting values for the severity of COVID-19 patients. *medRxiv* https://doi.org/10.1101/2020.05.01.20086421 (2020).
39. Nigro, P., Pompilio, G. & Capogrossi, M. C. Cyclophilin a: a key player for human disease. *Cell Death Dis.* **4**, e888 (2013).
40. Liu, X. *et al.* Cyclosporin a inhibits the influenza virus replication through cyclophilin a-dependent and -independent pathways. *PLoS ONE* **7**, e37277 (2012).
41. Liu, X. *et al.* Cyclophilin a restricts influenza a virus replication through degradation of the m1 protein. *PLoS ONE* **7**, e31063 (2012).
42. Kaul, A. *et al.* Essential role of cyclophilin a for hepatitis c virus replication and virus production and possible link to polyprotein cleavage kinetics. *PLoS Pathog.* **5**, e1000546 (2009).
43. Nakagawa, M. *et al.* Suppression of hepatitis c virus replication by cyclosporin a is mediated by blockade of cyclophilins. *Gastroenterology* **129**, 1031–1041 (2005).
44. Watashi, K., Hijikata, M., Hosaka, M., Yamaji, M. & Shimotohno, K. Cyclosporin a suppresses replication of hepatitis c virus genome in cultured hepatocytes. *Hepatology* **38**, 1282–1288 (2003).
45. Inoue, K. *et al.* Combined interferon alpha2b and cyclosporin a in the treatment of chronic hepatitis c: controlled trial. *J. Gastroenterol.* **38**, 567–572 (2003).
46. Flisiak, R. *et al.* The cyclophilin inhibitor debio 025 combined with peg ifnalpha2a significantly reduces viral load in treatment-naïve hepatitis c patients. *Hepatology* **49**, 1460–1488 (2009).
47. Gallay, P. A. *et al.* The novel cyclophilin inhibitor cpi-431-32 concurrently blocks hcv and hiv-1 infections via a similar mechanism of action. *PLoS ONE* **10**, e0134707 (2015).
48. Ptak, R. G. *et al.* Inhibition of human immunodeficiency virus type 1 replication in human cells by debio-025, a novel cyclophilin binding agent. *Antimicrob. Agents Chemother.* **52**, 1302–1317 (2008).
49. Pfefferle, S. *et al.* The sars-coronavirus-host interactome: Identification of cyclophilins as target for pan-coronavirus inhibitors. *PLoS Pathog.* **7**, e1002331 (2011).
50. Russell, C. D. & Haas, J. Cyclosporine has a potential role in the treatment of sars. *J. Infect.* **67**, 84–85 (2013).
51. Sun, J. *et al.* Comparative transcriptome analysis reveals the intensive early-stage responses of host cells to SARS-CoV-2 infection. *J. Infect.* https://doi.org/10.1101/2020.04.30.071274 (2020).
52. Satış, H. *et al.* Prognostic value of interleukin-18 and its association with other inflammatory markers and disease severity in covid-19. *Cytokine* **137**, 155302 (2020).
53. Gabay, C. *et al.* Open-label, multicentre, dose-escalating phase ii clinical trial on the safety and efficacy of tadekinig alfa (il-18BP) in adult-onset stills disease. *Ann. Rheum. Dis.* **77**, 840–847 (2018).
54. Krumm, B., Meng, X., Xiang, Y. & Deng, J. Identification of small molecule inhibitors of interleukin-18. *Sci. Rep.* **7**, 483 (2017).
55. Kolde, R., Laur, S., Adler, P. & Vilo, J. Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics* **28**, 573–580 (2012).
56. Hoffmann, M. *et al.* SARS-cov-2 cell entry depends on ace2 and tmprss2 and is blocked by a clinically proven protease inhibitor. *Cell* **181**, 271-280.e8 (2020).
57. Wan, Y., Shang, J., Graham, R., Baric, R. S. & Li, F. Receptor recognition by the novel coronavirus from wuhan: an analysis based on decade-long structural studies of sars coronavirus. *J. Virol.* **94**, 7 (2020).
58. Mick, E. *et al.* Upper airway gene expression differentiates covid-19 from other acute respiratory illnesses and reveals suppression of innate immune responses by sars-cov-2. *medRxiv* https://doi.org/10.1101/2020.05.18.20105171 (2020).

Acknowledgements

The authors would like to acknowledge Prof. C. Wiley and Prof J. Doench for generously sharing unpublished data and for their helpful comments on the manuscript. Work on this project was funded by UKRI (BBSRC and Medical Research Council [grant MC_PC_19059]).

Author contributions

N.P., N.R., M.H.F., C.D.R., B.W., M.Z., M.C.S., J.E.M. and S.C. collected, curated and processed included data. N.P., B.W. and S.C. completed the analyses. N.P., M.Z., M.C.S., and S.C. prepared the figures. B.W., A.L. and J.K.B. wrote and maintained the MAIC algorithm and scripts. N.P., N.R., M.H.F., C.D.R., J.K.B. and S.C. wrote the manuscript. C.D.R., J.K.B. and S.C. supervised the project. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-020-79033-3.

Correspondence and requests for materials should be addressed to J.K.B. or S.C.