Transfer transcriptomic signatures for infectious diseases

Julia di Iulio[a], Istvan Bartha[a], Roberto Spreafico[a], Herbert W. Virgin[a,b,c,1], and Amalio Telenti[a,1]

[a]Vir Biotechnology, Inc., San Francisco, CA 94158; [b]Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, MO 63110; and [c]Department of Internal Medicine, University of Texas Southwestern Medical Center, Dallas, TX 75390

Edited by Rafi Ahmed, Emory University, Atlanta, GA, and approved April 12, 2021 (received for review October 28, 2020)

The modulation of the transcriptome is among the earliest responses to infection. However, defining the transcriptomic signatures of disease is challenging because logistic, technical, and cost factors limit the size and representativeness of samples in clinical studies. These limitations lead to a poor performance of signatures when applied to new datasets. Although the study focuses on infection, the central hypothesis of the work is the generalization of sets of signatures across diseases. We use a machine learning approach to identify common elements in datasets and then test empirically whether they are informative about a second dataset from a disease or process distinct from the original dataset. We identify sets of genes, which we name transfer signatures, that are predictive across diverse datasets and/or species (e.g., rhesus to humans). We demonstrate the usefulness of transfer signatures in two use cases: the progression of latent to active tuberculosis and the severity of COVID-19 and influenza A H1N1 infection. This indicates that transfer signatures can be deployed in settings that lack disease-specific biomarkers. The broad significance of our work lies in the concept that a small set of archetypal human immunophenotypes, captured by transfer signatures, can explain a larger set of responses to diverse diseases.

transcriptomics | immunophenotype | infection | vaccination | transfer learning

Infection and vaccination trigger a robust transcriptome response in tissues or in blood. These perturbations occur in the setting of a preexisting immunophenotype in each individual characterized by a transcriptome that is regulated by genetics, the environment, the microbiome and virome, and prior infections (1–7). In the case of acute infections such as viral respiratory diseases, viral disease transmitted by arthropod vectors, or in chronic viral infections such as HIV, responses in peripheral blood mononucleolar cells may lead to the transcriptional deregulation of thousands of genes that vary significantly between individuals based on the status of their immune system at the time of infection (8–14). Cellular, biological, and functional factors contribute to this aggregate transcriptomic response. For example, changes in cell composition, the nature of inflammatory responses, bystander effects of tissue damage, and therapeutic intervention generate complex patterns of expression that vary between individuals. The diversity of approaches used to investigate transcriptome responses—study design, timing, technical platform—also contribute to the patterns of expression across studies. Additional sources of noise in transcriptome profiles may result from the relatively small sample size that characterize many publications and the pervasive impact of batch effects (15). Given these variations, it is perhaps not surprising that consensus transcriptomic signatures that reliably operate across studies have been challenging to identify. This is important for designing prospective analyses of the human immunophenotype and to taking advantage of the wealth of legacy data from earlier work and data repositories, for example, to establish host response–based diagnostics (9, 16), with an emphasis on meta-analytical approaches (17, 18).

Because of the above considerations, there is significant interest in developing methods that reproducibly identify transcriptome profiles as biomarkers of disease susceptibility or prediction of vaccine responses. A key challenge is the generation of appropriate datasets for each pathogen and study endpoint. This effort requires considerable planning and resourcing. Once biomarkers are identified, they still need to undergo extensive validation in additional cohorts and applied in settings other than the population in which the study was originally conducted. Ideally, biomarkers would be so robust that they could be transferred across studies and possibly across pathogens and species. For example, the severity of responses to viral respiratory infections could be predicted using a set of shared responses based on a strong deregulation of interferon-stimulated genes observed in the setting of different viral infections (19). Similarly, vaccine protection could be predicted using shared markers of response (12). Underlying these questions is the possibility of baseline human immunophenotypes that can be predictive of differential responses to various perturbations (20, 21). The overarching concept tested herein is that while the broad field of biomarkers—and specifically transcriptomics-based biomarkers—emphasizes specificity (to pathogen, perturbation, and/or study endpoint), we hypothesize that there are sets of common responses having the desirable properties of generalization and transferability.

Here, we identify patterns of gene response comprising transfer signatures that can be learned from deposited datasets and tested for predictive power in independent transcriptomes associated with clinical metadata. This work evaluates the performance of such transfer signatures across pathogens and studies, including the validity of transfer signatures learned from animal models for studies of human disease. We present two use cases of transfer signatures for infection. Our work establishes the validity of this approach and explores the nature of human immunophenotypes. If generally applicable in additional studies, the methods described

Significance

Human responses to infection include transcriptional changes shared across diverse pathogens. To capture these common patterns, we establish the concept of, and the method for, the identification of "transfer signatures": sets of genes defining human immunophenotypes. We demonstrate the usefulness of transfer signatures in two use cases: the progression of latent to active tuberculosis and the severity of viral respiratory infections.

Author contributions: A.T. designed research; J.d.I. performed research; J.d.I. contributed new reagents/analytic tools; J.d.I., I.B., R.S., and A.T. analyzed data; and J.d.I., H.W.V., and A.T. wrote the paper.

Competing interest statement: All authors are employees of, and hold stock or stock options in, Vir Biotechnology, Inc. H.W.V. is a founder of Casma Therapeutics and PierianDx, neither of which funded the research reported herein.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: svirgin@vir.bio or atelenti@vir.bio.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2022486118/-/DCSupplemental.

Published May 24, 2021.

here may lead to a rapid evolution of clinically and biologically relevant concepts in immunity and pathogenesis.

Results

Selection of Signatures and Datasets. We conceived this study under the assumption that the literature provides short lists of genes (here described as "signatures") that are predictive of study outcomes across different biological settings. The study design is shown in Fig. 1. Previously identified and published signatures—referred to as "literature signatures"—may or may not be associated with full access to raw data/sequencing counts. Thus, we collected 148 literature signatures to support exploratory analysis and selected raw data from 15 studies to train and test machine learning models (*Materials and Methods* and Dataset S1). The pairing of literature signatures, training, and test datasets are depicted in Fig. 1, and the analytical are steps detailed in *SI Appendix*, Fig. S1. By design, we built the study on RNA sequencing and microarray datasets. Many comparative studies have shown that their results are not always consistent. These inconsistencies notwithstanding, transforming expression levels from either technology into biologically relevant gene set enrichment scores significantly increases their correlation (22).

Our study focused on infectious diseases; thus, we obtained literature signatures from papers investigating 1) infection with dengue, severe acute respiratory syndrome coronavirus-1 (SARS-CoV-1), SARS-CoV-2, Middle East respiratory syndrome coronavirus (MERS-CoV), influenza A virus (IAV) H1N1, H5N1, H3N2, measles, and respiratory syncytial virus (referred to as "infection signatures," $n = 43$); 2) vaccine response to hepatitis B virus (HBV), IAV H1N1, H3N2 and/or influenza B virus, and against tuberculosis (TB) and simian immunodeficiency virus (referred to as "vaccine signatures," $n = 13$); and 3) progression to active TB (referred to as "TB signatures," $n = 20$). We also explored the information that is encoded in collections of signatures that characterize cell composition (referred to as "cell type signatures," $n = 22$) (23) and biological states assessed through hallmark gene sets of the Molecular Signatures Database (referred to as "hallmark signatures," MSigDB, https://www.gsea-msigdb.org/gsea/msigdb/collections.jsp, N = 50). The reason for including the latter two groups was to evaluate whether these literature signatures could be informative of the processes under study. For example, cell type composition in bulk measurements is critical for the final aggregate readout of sequencing and hallmark gene sets because they are a compact representation of biological processes/pathways and may be informative. Of note, literature signatures were not restricted to human studies, as some were sourced from studies in rhesus and cynomolgus macaques (Dataset S1).

Five publications provided training datasets (*SI Appendix*, Fig. S1): one study on dengue infection (24), one study on IAV H1N1 infection (25), one study on trivalent influenza vaccination [comprising two cohorts, one with males and one with females (26)], one study on HBV vaccination (27), and one study on TB vaccination in rhesus macaques (28). Their descriptions and sources are provided in Dataset S1. Of note, these studies contain multiple ($n = 14$) nonindependent datasets representing different time points. We expected this design to help understand the biology of transcriptome signatures and to allow the determination of the earliest time points with predictive power.

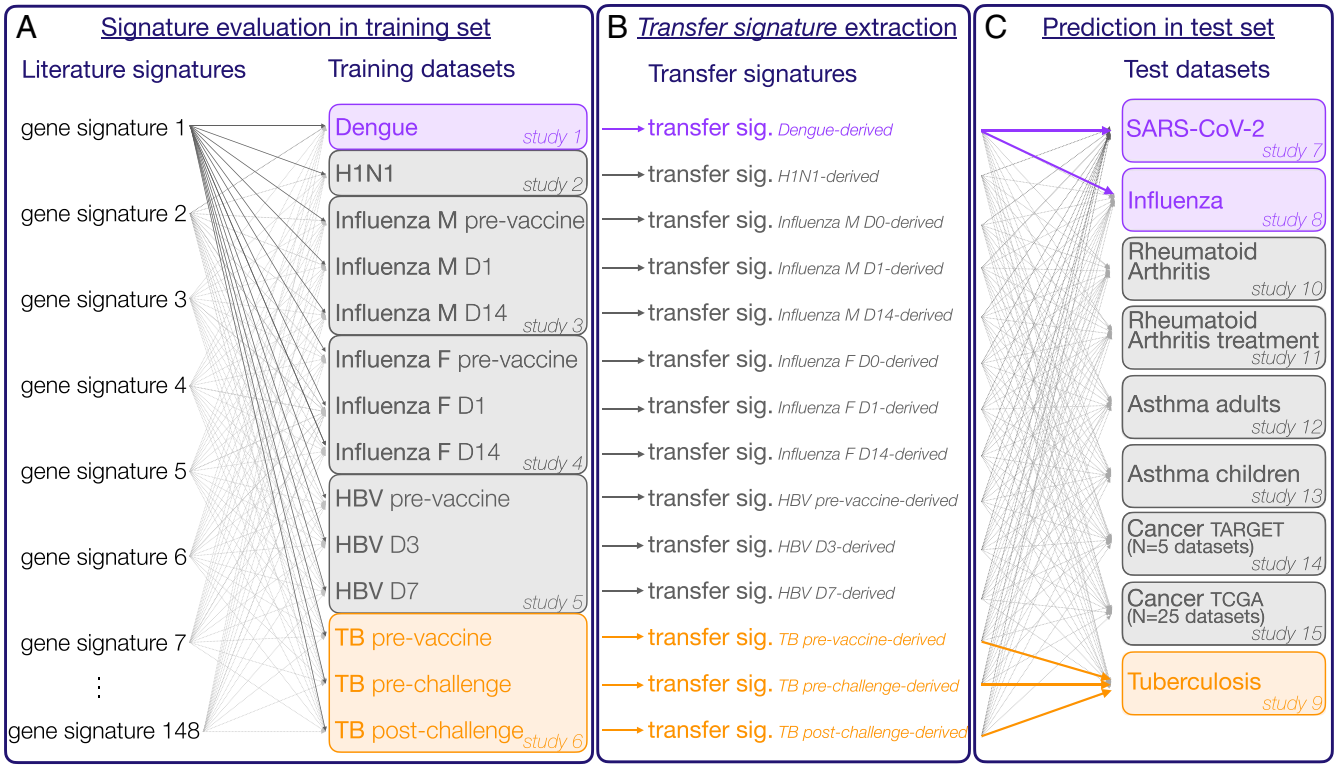

Fig. 1. Study design. Three steps to progress from (*A*) literature signatures to (*B*) transfer signatures to (*C*) prediction in unseen datasets. The study aims at predicting 1) SARS-CoV2 and influenza severe disease (purple) using a transfer signature extracted from a dengue infection dataset and 2) TB progression in humans (orange) using transfer signatures extracted from a Rhesus TB vaccine dataset. The study includes (in gray) other biologically related training datasets and other biologically related or unrelated test datasets to evaluate the performance of transfer signatures. For detailed information on the three steps, reference *SI Appendix*, Fig. S1. Description of signatures, datasets, and studies are provided in Dataset S1. D0, Day 0 is equivalent to prevaccine. D1, Day 1. D3, Day 3. D7, Day 7. D14, Day 14. F, Female. M, Male. TARGET, Therapeutically Applicable Research to Generate Effective Treatments. TCGA, The Cancer Genome Atlas.

Training and Testing of Transcriptome Signatures. We used random forest models to evaluate the collection of literature signatures on each training transcriptome dataset (Fig. 1*A*, *SI Appendix*, Fig. S1, and Dataset S1) followed by the extraction of a common set of predictive genes (transfer signature) from each training dataset (Fig. 1*B*, *SI Appendix*, Fig. S1, and Dataset S1). We then used the transfer signature obtained from one training dataset to predict the outcome in unseen unrelated test datasets using unsupervised

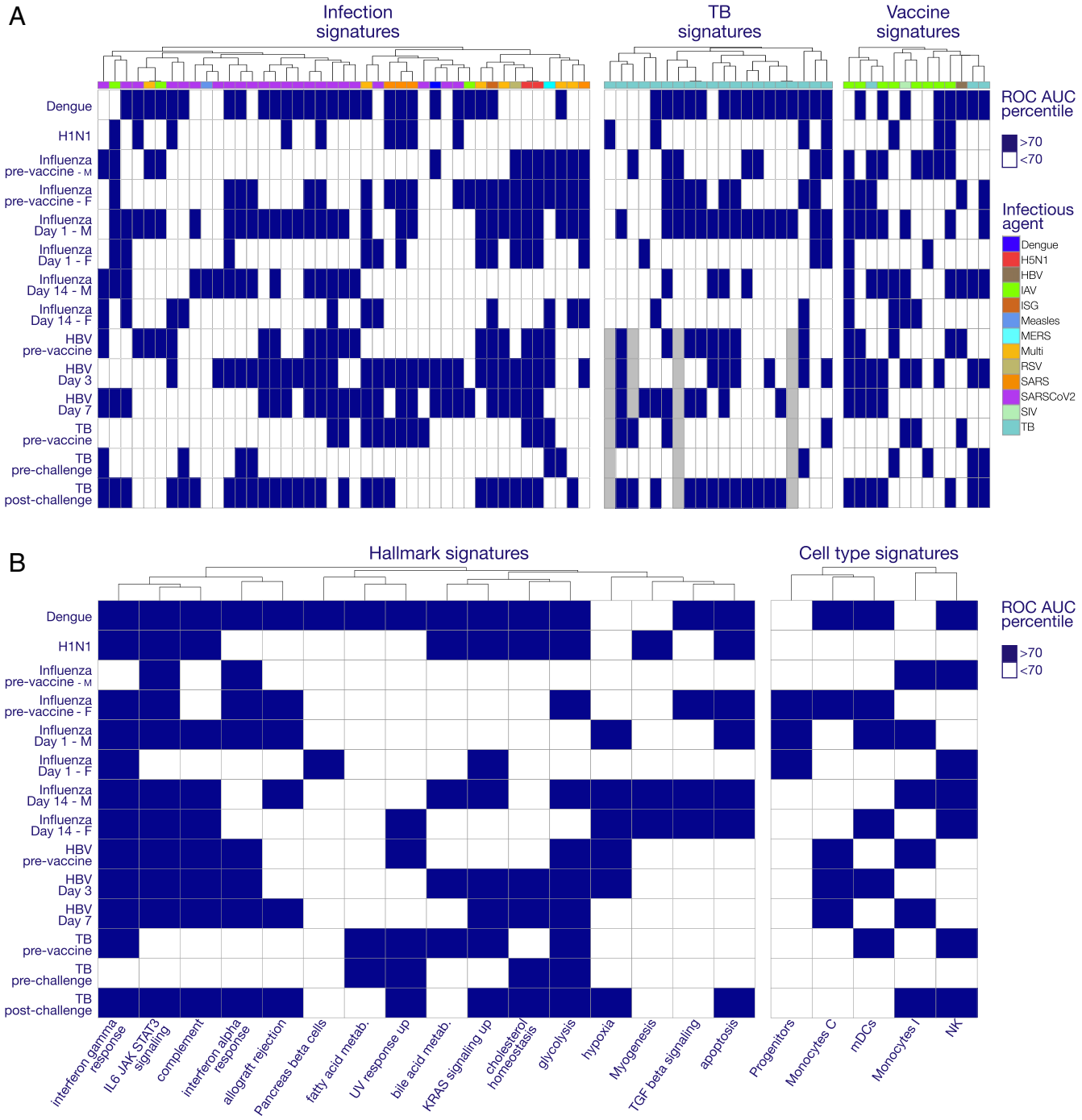

Fig. 2. Performance of literature signatures. (*A*) A heatmap of the AUROCs obtained through random forest models. Each column represents a signature from the literature grouped by signature group. Each row represents a training dataset. In order to be able to compare the AUROC across the datasets (which do not have the same case/control distribution), the AUROC are depicted in percentiles. The percentiles are obtained by comparing the performance of the literature signature to 100 random gene lists of the same size. The same cutoff as used for the signature retention in the model was used (70th percentile). Missing data are depicted in gray. The color annotation next indicates the infectious agent datasets. Influenza refers here to a trivalent vaccine consisting of H1N1, H3N2, and influenza B virus. (*B*) The best performing hallmark and cell type signatures. Each row represents a training dataset (in the same order as in *A*). Columns represent the signatures—hallmark (*Left*) and cell type (*Right*)—that reached the 70th percentile in at least one training dataset. For visual simplicity, the coloring here is binary as depicted in the legend. Metab., metabolism. IL, interleukin. JAK, Janus kinase. STAT, signal transducer and activator of transcription. mDC, myeloid dendritic cell. NK, natural killer cell. F, female. M, male. For additional information, see *Materials and Methods* and Dataset S1.

methods to exclude overfitting (Fig. 1*C*, *SI Appendix*, Fig. S1, and Dataset S1).

In the first step, in order to determine whether meaningful data exists in previously identified literature signatures for the prediction of orthogonal datasets, we characterized the performance of all 148 literature signatures on each training dataset. For each training dataset, we evaluated machine learning models with the feature set restricted to the genes contained in each previously identified literature signature. Effectively, for each of the 14 training datasets (for example, dengue infection), we obtained 148 models, yielding a total of 2,072 models across the training datasets. Then, for each model, we computed the receiver operating characteristic (ROC) values and the individual importance of each gene. We computed the ROC area under the curve (AUROC) using the leave-one-out cross validation strategy.

Because the percent split between cases and controls is different in each dataset, AUROCs could not be compared directly. We therefore expressed the results as percentiles rather than raw AUROC. The percentiles were obtained by comparing the performance of the literature signatures to random lists of genes of identical size as a control. We observed that a large proportion of literature signatures performed well across training datasets, supporting the notion that published signatures contain valuable shared information that can be used to train predictive models and classifiers (Fig. 2*A* and *SI Appendix*, Table S1). We compared the performance of the signature that was specifically provided in the original study of the training dataset against literature signatures (i.e., any other published signature considered in this study). All but one training dataset (HBV prevaccine) revealed at least one signature not reported in the original publication that outperformed the reported signature as measure by ROC or precision-recall (PR) AUCs (*SI Appendix*, Fig. S2). These results supported the concept that our approach may identify signatures that can be transferred between datasets while retaining predictive power.

Training and Testing of Hallmark and Cell Type Signatures. We also assessed two general subsets of literature signatures, one sourced from the Broad Institute's MSigDB hallmark pathway identifiers and one containing cell type signatures (23). These sets of literature signatures were not explicitly designed for association with infection diseases (Fig. 2*A*). Despite that, a number of these general-purpose literature signatures also performed well in predicting in several training datasets (*SI Appendix*, Table S1). As there is clear interest in understanding the nature of the signature that endows the generalization from hallmark and cell type signatures to other datasets, we further examined the signatures that performed well in at least one training dataset. Fig. 2*B* presents those top performing hallmark and cell type signatures across testing against our training datasets.

As several training datasets were time course experiments, we further inquired whether there existed patterns of performance across the hallmark and cell type signatures that would be consistent with the current understanding of biology. We noticed that the relevance of particular signatures shifted according to the timepoint in the experiment. For example, in the rhesus macaque TB vaccine experiment (Dataset S1, study 6), predictors of vaccine efficacy at the baseline included hallmark signatures of interferon gamma response, fatty acid metabolism, ultraviolet response, bile acid metabolism, KRAS signaling, and glycolysis as well as the cell type signatures of mDC and NK cells (Fig. 2*B*, TB prevaccine). In contrast, at the time of disease (TB postchallenge), the predictive signatures expanded to include hallmark IL6, JAK, STAT3 signaling, complement, interferon alpha response, allograft rejection, hypoxia, and apoptosis as well as a cell type signature of monocytes (Fig. 2*B*, TB postchallenge). The predictive value of these hallmark and cell type signatures is broadly consistent with the current understanding of biomarkers, pathogenesis, and cellular roles in TB (29–33). As another example, we observed differences in predictive signatures across gender category in the study of Franco et al. (26) that evaluated the response to influenza vaccination (Fig. 2). This is consistent with sex differences in the blood transcriptome associated with immune responses (34, 35).

Overall, the analysis of literature signatures from various sources supports the original hypothesis that there are shared response elements that serve as biomarkers across multiple conditions. On this basis, we next sought to create signatures with predictive power across a wider range of diseases (transfer signatures).

Defining Transfer Signatures. To establish a transfer signature for each training dataset, we used every literature signature that had

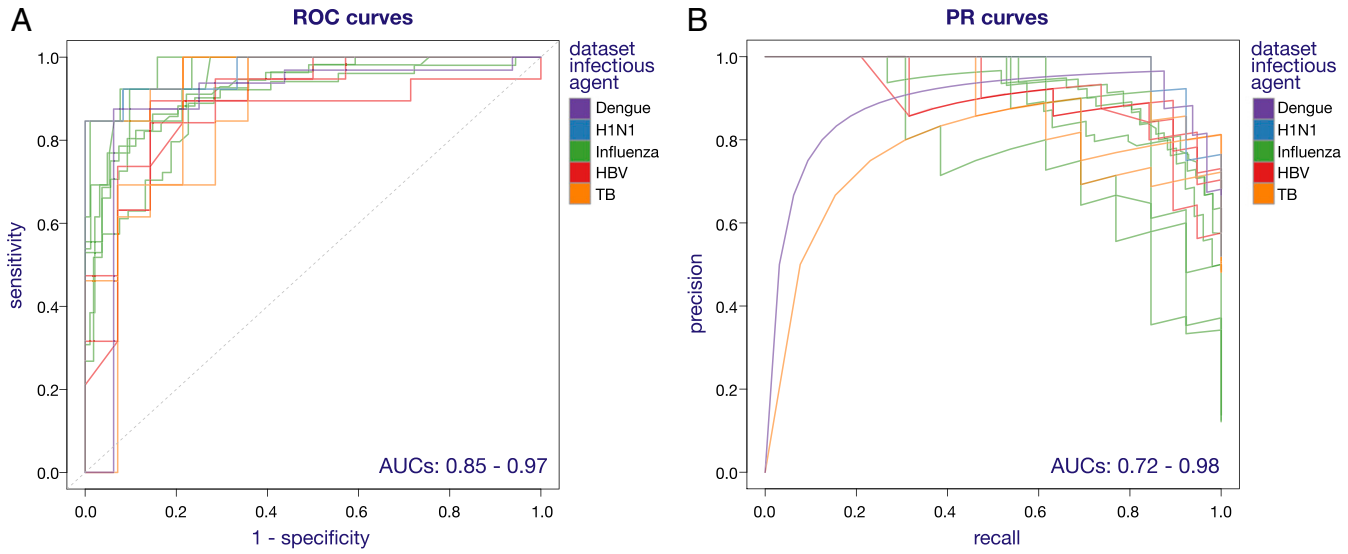

Fig. 3. Optimized performance of transfer signatures. The classifying performance of the predicted phenotypes obtained from the random forest models using the transfer signature was assessed for each respective training dataset—where the transfer signatures were obtained from *Materials and Methods* and Dataset S1. (*A*) ROC curves. (*B*) PR curves. Each line depicts the curve obtained for a given training dataset. The lines are colored based on the infectious agent studied in the training dataset.

a AUROC higher than the 70th percentile compared to random lists of genes of identical size (*Materials and Methods* and Fig. 1*B*). This step created one transfer signature per training dataset. Importantly, for the purpose of defining transfer signatures, we excluded the signature initially reported in each publication to minimize bias in the generation of transfer signatures and to allow an inclusion of genes that were potentially relevant in other published studies. This approach optimized the potential to detect signatures with generalizable properties. We then standardized the gene importance output from the random forest models of the signatures that passed the 70th percentile threshold and selected the first 50 genes (we also evaluated 10 and 20 gene signatures, *SI Appendix*, Fig. S3) with the highest standardized importance feature score (*Materials and Methods* and Fig. 1*B*). As expected—given that transfer signatures are made of the top-classifying genes for a given training dataset—transfer signatures performed well on the datasets they were trained on: AUROC varied between 0.85 and 0.97 and PR AUC of 0.72 to 0.98 for the various training datasets (Fig. 3 *A* and *B*). In all but one training dataset (TB prevaccine), transfer signatures matched or improved the performance, in terms of AUROC, of the best single performing literature signature. Furthermore, the transfer signatures outperformed the original signatures identified in each individual publication (*SI Appendix*, Fig. S4). The nature of the genes retained in the transfer signatures

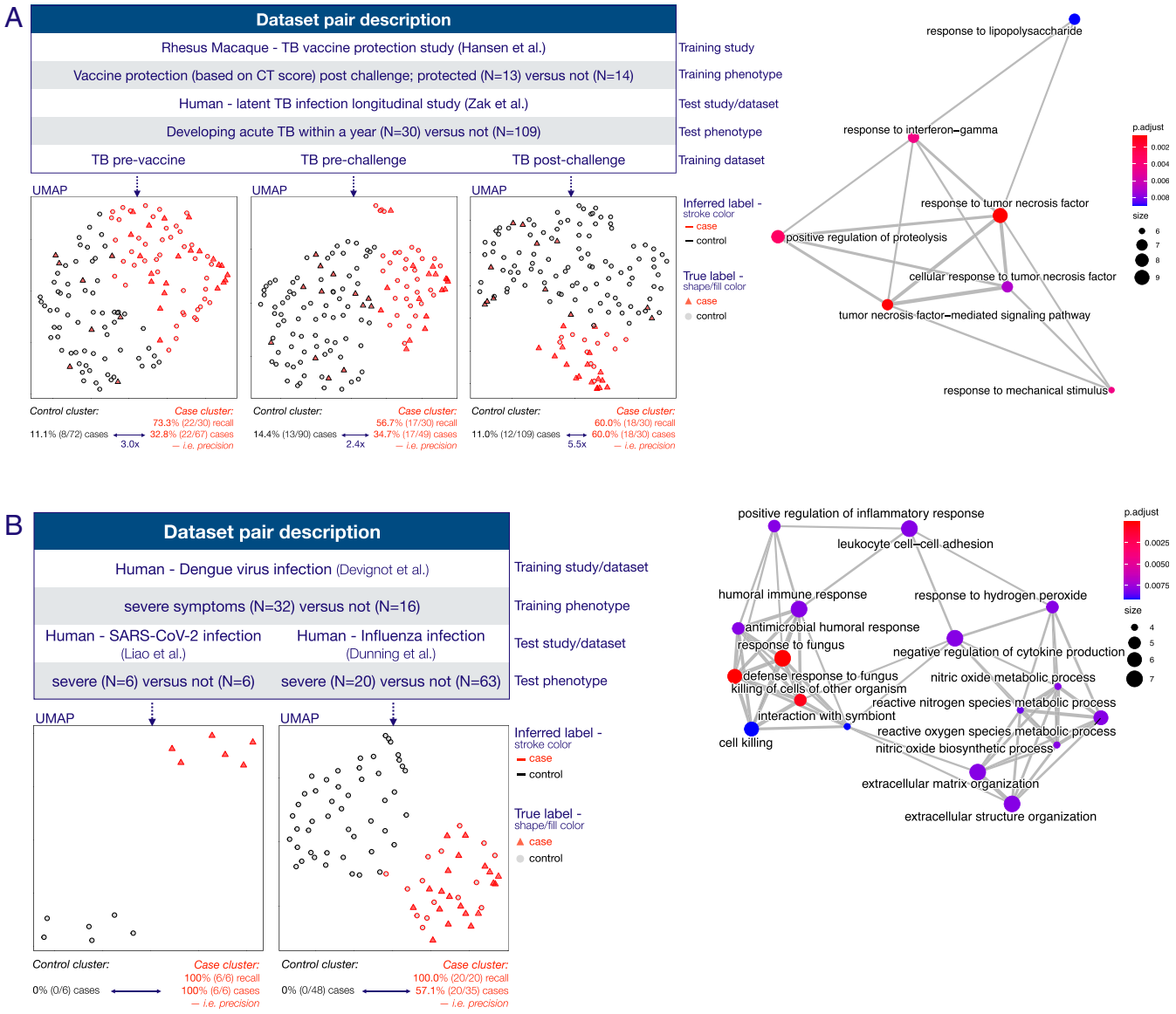

Fig. 4. TB and severe viral disease use cases—performance of unsupervised clustering. (*A*, *Top*) TB progression study design. (*A*, *Bottom*) Uniform Manifold Approximation and Projection (UMAP) of the test dataset using the 50-gene long transfer signature obtained from the respective training dataset shown in *Top*: prevaccine, preinfectious challenge, and postchallenge. (*B*, *Top*) Severe viral disease study design. (*B*, Bottom) UMAP projection of the test datasets using the 50-gene long transfer signature obtained from the training dataset shown in *Top*. For both panels, each sample is represented by round or triangle. The stroke color indicates the inferred label (from the unsupervised clustering), and the shape and fill color indicate the true label. The recall and percentage of true cases in the different clusters (i.e., precision) is displayed below each UMAP projection. For both panels, we summarize the biological content of the transfer signatures (TS) by displaying the gene set overrepresentation performed on the Biological Process Gene Ontology (GO) (*Top*: TB postchallenge TS; *Bottom*: dengue TS). Dots represent term enrichment with color coding: red indicates high enrichment, and blue indicates low enrichment. The sizes of the dots represent the percentage of contributing genes in a GO term. The significance was judged by a Benjamini–Hochberg correct *P* value cutoff of 0.01.

(listed in Dataset S1) prominently include immune system and metabolic processes.

By selecting genes for high importance from random forest models, transfer signatures become inevitably optimized and potentially overfit for each training study dataset. As we expect that overfitting will limit the generalizability of signatures to new datasets, we implement transfer signatures as lists of genes with no weights attached. This approach was undertaken so that errors from overfitting were not carried forward when using transfer signatures for prediction in unseen data.

Predictive Power of Transfer Signatures in Unseen Data. The use of a transfer signature in an unsupervised approach, that is, without retraining with known labels, is the most stringent implementation of the transfer signature. This is in particular true when the training dataset is chosen to be only partially related, as we currently understand biology, to the target test dataset.

In a first use case, we tested the hypothesis that information from a TB vaccine study in rhesus macaques can inform on progression from latent to active TB in a human cohort, that is, interspecies application of a transfer signature. In the second use case, we tested whether information from severe dengue infection serves for the classification of cases of severe SARS-CoV-2 and of influenza A infection, that is, interinfectious disease transfer signature application (Fig. 1 and Dataset S1). In these two scenarios, we expressed the outcome as "enrichment," that is, the ability of a transfer signature to increase the number of true cases in a population, and "recall," that is, the fraction of cases that are retrieved in a given subpopulation selected using the transfer signature.

Use case one: Progression of latent to active TB. We modeled the value of the transfer signature obtained in an animal study to assess the challenge of identifying a subpopulation of subjects within a clinical trial that are likely to reach a given clinical endpoint. The scenario is the use of a pharmacological or vaccine intervention to prevent progression from latent TB to active disease. Progression to active TB is a rare event (estimated as 0.084 cases per 100 person/years) (36); therefore, it would be important to be able to recruit individuals that are the most likely to develop active infection within 1 y. Indeed, in the presence of a limited numbers of individuals that may reach a specific endpoint, the study may lack power to detect differences between the placebo and vaccine or treatment group.

Here, we tested transfer signatures obtained with the three time course datasets from Hansen et al. (28) (Dataset S1, study 6). Effectively, this implies training of all literature signatures (excluding the signature identified in the original publication) on each of the three datasets, selecting the best performing genes for each respective dataset (Fig. 1). This training data derived from a study that assessed the efficacy of a TB vaccine in rhesus macaques, with longitudinal samples from 27 rhesus macaques collected prevaccine, after vaccination, but before TB challenge and 4 wk postchallenge. The phenotype used for training the random forest models was vaccine efficacy in protection from TB defined as a computed tomography score of <10 (protected, $n = 13$) at any time point postchallenge versus not protected ($n = 14$).

We used as a target dataset the data from Zak et al. (37), a longitudinal study assessing progression from latent to active TB (Dataset S1, study 9). We defined cases as individuals that developed TB within a year ($n = 30$) and controls as individuals that did not develop TB within a year after entry in the study ($n = 109$; Dataset S1). The results of the unsupervised clustering are shown in Fig. 4*A*. With the transfer signature defined on the prevaccine rhesus macaque samples, 32.8% (22/67) of the predicted cases were true cases, that is, developed active TB within a year, while the samples outside of this cluster contained only 11.1% (8/72) of true cases. Here, the unsupervised clustering led to a threefold enrichment (when comparing cases versus noncase cluster or a 1.5-fold enrichment when comparing the case cluster versus the general population) and a 73.3% recall. In a similar setting, but with the transfer signature derived from postvaccination but prechallenge samples from macaques, we obtained a twofold enrichment (34.7% [17/49] versus 14.4% [13/90] when comparing cases versus noncase cluster or a 1.6-fold enrichment when comparing the case cluster versus the general population) and a 56.7% recall. With the transfer signature derived from postchallenge samples, we obtained a 5.5-fold enrichment (60.0% [18/30] versus 11.0% [12/109] when comparing cases versus noncase cluster or 2.8-fold enrichment when comparing the case cluster versus the general population) and 60.0% recall. Analysis of the content of the TB transfer signatures confirmed the enrichment of genes of the immune response (Dataset S1); we display the gene set overrepresentation for the most predictive TB transfer signature in Fig. 4*A*. Overall, the use of the transfer signatures from this animal model would enable the prospective recruitment of individuals into smaller clinical trials with a greater likelihood of reaching adequate end point events to allow statistical power.

Use case two: Severity of viral infection. We next assessed whether transfer signatures could be used in the setting of viral infection to predict the severity of the symptoms of individuals that are hospitalized. Here, we tested transfer signatures obtained from the dataset from Devignot et al. (24) (Dataset S1, study 1), defining transcriptomes of children with acute dengue infection whose blood samples were collected within 3 to 7 d of the onset of fever. For our analysis, we considered as cases the children with severe manifestations of disease (shock syndrome and hemorrhagic fever; $n = 32$), while children that had uncomplicated dengue fever were considered controls ($n = 16$). We then used data from Liao et al. (38) (Dataset S1, study 7) and Dunning et al. (8) (Dataset S1, study 8) as two different target datasets. The phenotypes in these studies were established at the time when, or before, the RNA samples were obtained. Therefore, the unsupervised clustering results (Fig. 4*B*) reflect here the performance of transfer signatures as classifiers rather than predictors.

The study of Liao et al. (38) characterized bronchoalveolar lavage fluid immune cells from patients infected with SARS-CoV-2. For the purpose of this analysis, we considered as cases the individuals that were described in the original report as having severe disease ($n = 6$), while individuals with moderate disease ($n = 3$) or not infected ($n = 3$) were considered as controls (total $n = 6$). The RNA samples were obtained 4 to 10 d after the phenotypes were established. Using a transfer signature derived from transcriptomes of children with severe dengue, all true cases of severe SARS-CoV-2 were correctly classified. This represented 100% precision and 100% recall with a twofold enrichment when comparing the case cluster versus the total population studied (Fig. 4 *B*, *Left Lower*).

The study of Dunning et al. (8) characterized blood samples from individuals hospitalized with influenza. We considered as cases the individuals that were described in the original report as requiring mechanical ventilation ($n = 20$), while individuals that did not require mechanical ventilation were considered as controls ($n = 63$). The use of a transfer signature from children with severe dengue allowed us to infer a case cluster that included 57.1% of the severe cases, while none of the severe cases appeared in the inferred control cluster. This corresponds to a 2.4-fold enrichment of severe cases when comparing the case cluster versus the total population studied and a 100% recall (Fig. 4 *B*, *Right Lower*). Analysis of the content of the dengue transfer signature confirmed the enrichment of genes of the immune response (Dataset S1 and Fig. 4*A*).

Of note, the results displayed for the various use cases used a 50-gene long transfer signature; however, similar results were obtained when selecting only the top 20 genes, while the performance dropped with some of the 10-gene transfer signatures (*SI Appendix*, Figs. S5 and S6). We obtained similar results when using transfer signatures derived with only hallmark signatures (Dataset S1) compared to transfer signatures based on all literature signatures (*SI Appendix*, Figs. S7 and S8). Overall, both the

SARS-CoV-2 and the influenza studies support the value of the transfer of signatures, as defined by our approach, across different viral infections to classify disease severity.

Choice of training datasets for unseen data. The successful implementation of transfer signatures described above leaves open the question of how to choose the optimal transfer signature to be applied in a new dataset. We approached this question with the parsimonious initial approach of selecting training and test datasets from diseases that, based on our understanding of disease pathogenesis, might be related. For example, TB vaccination efficacy might reasonably be assumed to relate to the prevention of the progression of TB, and the severity of viral disease caused by dengue, SARS-CoV-2, and influenza might reasonably be considered to be related. To challenge this biological understanding–biased decision, we next evaluated the performance of transfer signatures and test datasets from biological processes that were less clearly related. To this end, we chose to test the transfer signatures above and additional transfer signatures from influenza and hepatitis B vaccination (Fig. 1*B*) to predict the severity of inflammatory and autoimmune diseases (rheumatoid arthritis and asthma) and to predict survival from malignancy as measured in datasets from cancer (Fig. 1*C*).

As hypothesized, the original training test pairs from diseases with more apparent biological relationships (dengue and SARS-CoV-2 and influenza; TB in an animal model and in humans) were optimal choices ("related pairs," Dataset S2). However, we also observed good performance for severe respiratory viral infection transfer signatures in rheumatoid arthritis (Dataset S2), which reinforces the concept of shared immunophenotypes and suggests that diseases with less apparent relationships clinically may nevertheless have underlying similarities in biology that are identified by our machine learning–based approach. In addition, some transfer signatures were occasionally predictors of outcome for certain cancer types ("unrelated pairs," Dataset S2). Although these observations may be expected by chance, it is worth discussing some of the salient instances of transfer signature enrichment in cancer. A prognosis of uterine carcinosarcoma and Wilms tumor were associated with the transfer signature of a severe respiratory viral infection—a finding potentially related to the previously identified inflammatory microenvironment of these cancers (39, 40). A prognosis of acute myeloid leukemia was associated with a TB transfer signature, a finding potentially related to the fact that hematologic malignancies can be accompanied by the overproduction of inflammatory cytokines based on different cellular origins and concurrent chronic inflammatory responses (41). These observations extend the interest of exploring transfer signatures from infectious diseases to unrelated fields such as autoimmunity and in cancer.

Discussion

The present work considers a transcriptomic response as a complex set of expression profiles that includes genes uniquely modulated by a given perturbation as well as shared responses (e.g., up-regulation of interferon-stimulated genes), cell type composition, prior history of infection, and influence of genetics and the environment on the individual as well as experimental noise. A single study will inevitably sample from a complex transcriptome response to generate a study-specific signature. Because many studies are limited in size, the resulting signatures will be, to some extent, the result of a random sampling of the large space of deregulated genes and of noise. Against this backdrop, a transfer signature aims at capturing informative markers of the broadest use. Consistent with our working hypotheses, we found that many signatures derived from data in the literature are informative when tested in other datasets, including for related studies (e.g., same pathogen) or even across pathogens, vaccines, and phenotypes. Recognizing this, we systematized the identification of transfer signatures using the concepts of transfer learning in the field of artificial intelligence, wherein a model trained to solve one task can be repurposed on a related task. Transfer learning also uses information from larger compendia of datasets to inform and constrain the models, and our work included this principle as well via inclusion of pathway and cell type signatures. Our work and that of others also supports the possibility of extending prediction across species; for example, a recent study indicates that the mouse transcriptome reveals potential signatures of protection and pathogenesis in human TB (42).

The concept of transferability applies to two distinct steps in the analysis presented here. In a first step, gene lists from a multiplicity of unrelated datasets (termed literature signatures) are applied to new training datasets. This step identified a good performance of literature signatures over random gene lists of the same length. In a second step, a gene list generated by machine learning from the ensemble of literature signatures through the use of a training dataset—a transfer signature—is tested on an unseen dataset to assess performance. This second step examines the optimal predictive and classifying power of transfer signatures.

Our approach also revealed significant information content in more general sets of initial signatures, such as those listed in the hallmark gene sets of the MSigDB, a result consistent with recent observations (43). We interpret this as an indication that even in the absence of preexisting information in the literature, well-defined sets of hallmark genes will enable the extraction and creation of transfer signatures. Similarly, we observed that signatures of cell type composition (23) have also classifying power and could possibly inform on mechanisms of pathogenesis. A number of resources use coexpression patterns to classify global transcriptome patterns (44, 45) or to allow deconvolution of cellular content (46). In general, most gene sets in those resources are curated pathways (e.g., most of MSigDB). In addition, most of the time, functional enrichment is done against the pathway gene sets, not against experimental gene sets. Our approach goes one step forward in 1) generalizing the concept to reuse signatures/families of signatures and 2) creating a machine learning infrastructure to define the most discriminative learned signatures for a particular disease/set of diseases, that is, transfer signatures.

We describe two use cases that illustrate the value of transfer signatures in predicting the progression of TB and in classifying the severity of viral respiratory disease. We suggest that transfer signatures may therefore serve as biomarkers in clinical medicine. For example, there are field applications of transcriptome signatures created through metanalysis of up to 16 different studies for the prediction of the progression of latent TB to active disease (16) and from 17 studies to identify signatures for incipient TB (47). The capability to identify such subjects is paramount to targeting treatment of latent TB to individuals at the highest risk for progression. We also emphasize the strategy of using transfer signatures to enrich clinical trials for subjects that have a greater likelihood of developing an endpoint. Such enriched clinical trial designs could be smaller and, thus, cheaper.

It is possible that collections of validated transfer signatures could serve as a basis to explore human immunophenotypes: the baseline conditions that associate with the notable diversity of individual responses in human when exposed to, for example, infection or vaccines (48). The definition of human immunophenotypes currently requires complex genetic and immunological phenotyping (20, 21). In the use cases, a number of the transfer signatures were trained in samples at "baseline"—for example, after vaccination, before the development of disease—and still generated good classification power or a prediction of distant endpoints. Thus, our work defines a general approach that creates transfer signatures to support immunophenotyping. A related concept is the interesting scenario of vaccine repurposing that has been recently discussed in the context of vaccines for SARS-CoV-2 infection (49). This concept implies that the general or broad responses to a given stimulus, here an unrelated

antigen, may confer protection to a second pathogen by means of common or shared responses.

Further work in very large-scale datasets collected prospectively and that contain relevant clinical metadata may allow the identification of parameters that support or render irrelevant the transfer learning approach. We speculate that such factors may include the identification of diseases states that do or do not have any underlying commonality. We predict that data from diseases with an inflammatory component (many diseases) may fall into different immunophenotype groupings, while noninflammatory diseases will not segregate with transfer signatures containing inflammation-related genes. It is possible, in contrast, that transfer signatures from seemingly unrelated disease will be shared. We argue that this is a significant value of our approach, as this would generate hypotheses by which seemingly unrelated diseases are related by underlying mechanisms. For example, we observed the potential value of transfer signatures from infectious diseases to serve as predictors of outcome for autoimmune diseases and for a subset of cancers—an important area for future research. This might have the effect of leveraging the depth of understanding of some diseases to explore less well-understood diseases and rare diseases that are not easily approached on a population basis. This could yield drugs or clinical interventions that might be applied to poorly understood or understudied diseases. A final goal is to define how many discrete and distinct transfer signatures/immunophenotypes can be defined and whether there are combinations of such elemental immunophenotypes. In summary, using machine learning approaches, we established the feasibility to transferring optimized gene shortlists from multiple studies to a target study with the retention of predictive and classifier power. This could significantly facilitate the use of transfer signatures for prospective studies before disease- and case-specific signatures can be determined.

Materials and Methods

SI Appendix contains a methods section that describes in detail the datasets and experimental procedures used in this study, including the signature descriptions, sources, references, and gene lists as well as the training and test datasets. We used categorical/binary phenotypes in order to be consistent across datasets. A random forest model was run on each "literature signature–training dataset" pair. The models were trained using the leave-one-out cross validation. The ROC and PR AUC were computed using the scores of the single left-out sample per trained model. For the extraction of transfer signatures, we used literature signatures that had a ROC AUC percentile above a defined threshold. Transfer signatures of length $n = 10$, 20, and 50 genes were tested empirically. Transfer signatures were used in an unsupervised analysis to cluster samples from independent test datasets. Dimension reduction was performed using Uniform Manifold Approximation and Projection followed by Hierarchical Density-Based Spatial Clustering.

Data Availability. All study data are included in the article and/or supporting information (URLs of the datasets are provided in Dataset S1). Code to reproduce this work is available in GitHub (https://github.com/virbio/manuscript-transfer-signatures).

ACKNOWLEDGMENTS. We thank C. Maher and X. Ding for useful commentaries. This research is funded by Vir Biotechnology, Inc.

1. P. Brodin *et al.*, Variation in the human immune system is largely driven by non-heritable influences. *Cell* **160**, 37–47 (2015).
2. L. K. Beura *et al.*, Normalizing the environment recapitulates adult human immune traits in laboratory mice. *Nature* **532**, 512–516 (2016).
3. E. S. Barton *et al.*, Herpesvirus latency confers symbiotic protection from bacterial infection. *Nature* **447**, 326–329 (2007).
4. T. A. Reese *et al.*, Sequential infection with common pathogens promotes human-like immune gene expression and altered vaccine response. *Cell Host Microbe* **19**, 713–719 (2016).
5. H. W. Virgin, The virome in mammalian physiology and disease. *Cell* **157**, 142–150 (2014).
6. T. S. Stappenbeck, H. W. Virgin, Accounting for reciprocal host-microbiome interactions in experimental science. *Nature* **534**, 191–199 (2016).
7. C. Moon *et al.*, Vertically transmitted faecal IgA levels determine extra-chromosomal phenotypic variation. *Nature* **521**, 90–93 (2015).
8. J. Dunning *et al.*; MOSAIC Investigators, Progression of whole-blood transcriptional signatures from interferon-induced to neutrophil-associated patterns in severe influenza. *Nat. Immunol.* **19**, 625–635 (2018).
9. M. Robinson *et al.*, A 20-gene set predictive of progression to severe dengue. *Cell Rep.* **26**, 1104–1111.e4 (2019).
10. R. Barral-Arca *et al.*, A meta-analysis of multiple whole blood gene expression data unveils a diagnostic host-response transcript signature for respiratory syncytial virus. *Int. J. Mol. Sci.* **21**, 1831 (2020).
11. H. I. Nakaya *et al.*, Systems biology of vaccination for seasonal influenza in humans. *Nat. Immunol.* **12**, 786–795 (2011).
12. S. Li *et al.*, Molecular signatures of antibody responses derived from a systems biology study of five human vaccines. *Nat. Immunol.* **15**, 195–204 (2014).
13. M. Rotger *et al.*, Comparative transcriptomics of extreme phenotypes of human HIV-1 infection and SIV infection in sooty mangabey and rhesus macaque. *J. Clin. Invest.* **121**, 2391–2400 (2011).
14. M. Rotger *et al.*; Swiss HIV Cohort Study; Center for HIV/AIDS Vaccine Immunology, Genome-wide mRNA expression correlates of viral control in CD4+ T-cells from HIV-1-infected individuals. *PLoS Pathog.* **6**, e1000781 (2010).
15. J. T. Leek *et al.*, Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* **11**, 733–739 (2010).
16. H. Warsinske, R. Vashisht, P. Khatri, Host-response-based gene signatures for tuberculosis diagnosis: A systematic comparison of 16 signatures. *PLoS Med.* **16**, e1002786 (2019).
17. T. E. Sweeney, W. A. Haynes, F. Vallania, J. P. Ioannidis, P. Khatri, Methods to increase reproducibility in differential gene expression via meta-analysis. *Nucleic Acids Res.* **45**, e1 (2017).
18. S. Blankley *et al.*, A 380-gene meta-signature of active tuberculosis compared with healthy controls. *Eur. Respir. J.* **47**, 1873–1876 (2016).
19. E. V. Mesev, R. A. LeDesma, A. Ploss, Decoding type I and III interferon signalling during viral infection. *Nat. Microbiol.* **4**, 914–924 (2019).
20. A. Alpert *et al.*, A clinically meaningful metric of immune age derived from high-dimensional longitudinal monitoring. *Nat. Med.* **25**, 487–495 (2019).
21. E. Patin *et al.*; Milieu Intérieur Consortium, Publisher Correction: Natural variation in the parameters of innate immune cells is preferentially driven by genetic factors. *Nat. Immunol.* **19**, 645 (2018).
22. F. M. van der Kloet, J. Buurmans, M. J. Jonker, A. K. Smilde, J. A. Westerhuis, Increased comparability between RNA-Seq and microarray data by utilization of gene sets. *PLOS Comput. Biol.* **16**, e1008295 (2020).
23. G. Monaco *et al.*, RNA-seq signatures normalized by mRNA abundance allow absolute deconvolution of human immune cell types. *Cell Rep* **26**, 1627–1640.e7 (2019).
24. S. Devignot *et al.*, Genome-wide expression profiling deciphers host responses altered during dengue shock syndrome and reveals the role of innate immunity in severe dengue. *PLoS One* **5**, e11671 (2010).
25. J. F. Bermejo-Martin *et al.*, Host adaptive immunity deficiency in severe pandemic influenza. *Crit. Care* **14**, R167 (2010).
26. L. M. Franco *et al.*, Integrative genomic analysis of the human immune response to influenza vaccination. *eLife* **2**, e00299 (2013).
27. E. Bartholomeus *et al.*, Transcriptome profiling in blood before and after hepatitis B vaccination shows significant differences in gene expression between responders and non-responders. *Vaccine* **36**, 6282–6289 (2018).
28. S. G. Hansen *et al.*, Prevention of tuberculosis in rhesus macaques by a cytomegalovirus-based vaccine. *Nat. Med.* **24**, 130–143 (2018).
29. A. N. Martinez, S. Mehra, D. Kaushal, Role of interleukin 6 in innate immunity to Mycobacterium tuberculosis infection. *J. Infect. Dis.* **207**, 1253–1261 (2013).
30. Y. Gao *et al.*, STAT3 expression by myeloid cells is detrimental for the T- cell-mediated control of infection with Mycobacterium tuberculosis. *PLoS Pathog.* **14**, e1006809 (2018).
31. H. Esmail *et al.*, Complement pathway gene activation and rising circulating immune complexes characterize early disease in HIV-associated tuberculosis. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E964–E973 (2018).
32. P. Sampath, K. Moideen, U. D. Ranganathan, R. Bethunaickan, Monocyte subsets: Phenotypes and function in tuberculosis infection. *Front. Immunol.* **9**, 1726 (2018).
33. R. Roy Chowdhury *et al.*, A multi-cohort study of the immune factors associated with M. tuberculosis infection outcomes. *Nature* **560**, 644–648 (2018).
34. E. Bongen *et al.*, Sex differences in the blood transcriptome identify robust changes in immune cell proportions with aging and influenza infection. *Cell Rep.* **29**, 1961–1973.e4 (2019).
35. B. Piasecka *et al.*; Milieu Intérieur Consortium, Distinctive roles of age, sex, and genetics in shaping transcriptional variation of human immune responses to microbial challenges. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E488–E497 (2018).
36. K. M. Shea, J. S. Kammerer, C. A. Winston, T. R. Navin, C. R. Horsburgh Jr, Estimated rate of reactivation of latent tuberculosis infection in the United States, overall and by population subgroup. *Am. J. Epidemiol.* **179**, 216–225 (2014).
37. D. E. Zak *et al.*; ACS and GC6-74 cohort study groups, A blood RNA signature for tuberculosis disease risk: A prospective cohort study. *Lancet* **387**, 2312–2322 (2016).
38. M. Liao *et al.*, Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. *Nat. Med.* **26**, 842–844 (2020).

39. F. Guo, Y. Dong, Q. Tan, J. Kong, B. Yu, Tissue infiltrating immune cells as prognostic biomarkers in endometrial cancer: A meta-analysis. *Dis. Markers* **2020**, 1805764 (2020).
40. P. Maturu, "The inflammatory microenvironment in Wilms tumors" in *Wilms Tumor*, M. M. van den Heuvel-Eibrink, Ed. (Codon Publications, Brisbane, Australia, 2016), pp. 189–207.
41. B. M. Craver, K. El Alaoui, R. M. Scherber, A. G. Fleischman, The critical role of inflammation in the pathogenesis and progression of myeloid malignancies. *Cancers (Basel)* **10**, 104 (2018).
42. L. Moreira-Teixeira *et al*., Mouse transcriptome reveals potential signatures of protection and pathogenesis in human tuberculosis. *Nat. Immunol.* **21**, 464–476 (2020).
43. C. A. Targonski, C. A. Shearer, B. T. Shealy, M. C. Smith, F. A. Feltus, Uncovering biomarker genes with enriched classification potential from Hallmark gene sets. *Sci. Rep.* **9**, 9747 (2019).
44. D. Chaussabel, N. Baldwin, Democratizing systems immunology with modular transcriptional repertoire analyses. *Nat. Rev. Immunol.* **14**, 271–280 (2014).
45. M. C. Altman *et al*., A novel repertoire of blood transcriptome modules based on co-expression patterns across sixteen disease and physiological states [Preprint] (2019). https://www.biorxiv.org/content/10.1101/525709v1 (Accessed 10 May 2021).
46. F. Avila Cobos, J. Vandesompele, P. Mestdagh, K. De Preter, Computational deconvolution of transcriptomics data from mixed cell populations. *Bioinformatics* **34**, 1969–1979 (2018).
47. R. K. Gupta *et al*., Concise whole blood transcriptional signatures for incipient tuberculosis: A systematic review and patient-level pooled meta-analysis. *Lancet Respir. Med.* **8**, 395–406 (2020).
48. J. S. Tsang *et al*., Improving vaccine-induced immunity: Can baseline predict outcome? *Trends Immunol.* **41**, 457–465 (2020).
49. K. Chumakov, C. S. Benn, P. Aaby, S. Kottilil, R. Gallo, Can existing live vaccines prevent COVID-19? *Science* **368**, 1187–1188 (2020).