Supplementary Materials for

Transfer transcriptomic signatures for infectious diseases

Authors: Julia di Iulio, Istvan Bartha, Roberto Spreafico, Herbert W. Virgin, Amalio Telenti*
Correspondence to: atelenti@vir.bio

This PDF file includes:

Materials and Methods
Figs. S1 to S9

Other Supplementary Materials for this manuscript include the following:

Dataset S1
Dataset S2
Table S1

Materials and Methods

Gene signatures. We used five categories of signatures from publications, referred as "*literature signatures*": (i) curated sets of gene lists –referred as *hallmark signatures* (N=50, https://www.gsea-msigdb.org/gsea/msigdb/collections.jsp) (1), (ii) gene signatures associated with cell composition in PBMC –referred as *cell type signatures* (N=22) (2), (iii) vaccine protection and response signatures –referred as *vaccine signatures* (N=13), (iv) progression from latent to active TB infection signatures –referred as *TB signatures* (N=20) and (v) viral and bacterial infection signatures –referred as *infection signatures* (N=43). All signature descriptions, sources, references and gene lists are provided in **Dataset S1** and compiled in part in ref (3). Of note, due to gene nomenclature conversion issues, some signatures may be missing some genes identified in the parent paper.

Training datasets. We used 14 different training datasets from six studies: one study on dengue infection (4) (**Dataset S1** – study 1), one study on influenza H1N1 infection (5) (**Dataset S1** – study 2), one study on trivalent Influenza vaccination comprising two cohorts, one with males (**Dataset S1** – study 3) and one with females (6) (**Dataset S1** – study 4) – each comprising 3 datasets obtained at different timepoints (pre-vaccination, day 1 and day 14 post-vaccination), one study on hepatitis B virus (HBV) vaccination (7) (**Dataset S1** – study 5) – comprising 3 datasets obtained at different timepoints (pre-vaccination, day 3 and day 7 post-vaccination) and one study on tuberculosis (TB) vaccination in rhesus macaques (8) (**Dataset S1** – study 6) – comprising 3 datasets obtained at different timepoints (pre-vaccination, pre-challenge with TB and 28 days post-challenge with TB). Of note, several studies contain multiple non-independent datasets (or timepoints). This design is expected to help understand the biology of transfer transcriptome signatures and to monitor what are the earliest time points with predictive power.

Test datasets. We used 3 test datasets from three studies: one study on bronchoalveolar lavage in SARS-CoV-2 infection (9) (**Dataset S1** – study 7), one study on influenza infection (10) (**Dataset S1** – study 8) and one longitudinal study on TB progression in latently infected individuals (11) (**Dataset S1** – study 9). Of note, all test datasets were independent from each other and from any training datasets.

Phenotypes used. We explored multiple phenotypes in the training and test datasets, that can be categorized in four groups, namely (i) severity of symptoms during viral infection (for dengue, influenza and SARS-CoV-2 infection studies), (ii) vaccine response (for both HBV and influenza vaccination studies), (iii) disease state - for TB vaccination study in rhesus macaque, and (iv) time to disease in the longitudinal study TB progression. Further description and the number of individuals in each phenotype category per study are provided in **Dataset S1**. Of note, the phenotype extracted from the publicly available datasets is not necessarily the one used in the original study. The differences in phenotype definition, if any, are provided in **Dataset S1**. As an example, we used categorical/binary phenotypes even when the original study used numerical phenotype in order to be consistent across datasets and to better mimic future potential practical use cases.

Gene Signature evaluation in training datasets. A random forest model was run on each "literature signature – training dataset" pair (hereafter referred as S-D pair). In order to prevent overfitting the model to a specific pair and given the downstream goal of identifying genes that were common biomarkers across experiments and conditions, rather than specific to a single study or pair, hyperparameters were not tuned and were used as follow: number of trees (N=1,000); all other hyperparameters were the default in *randomForest* function from the R package "randomForest" (https://cran.r-project.org/web/packages/randomForest/index.html). In the model, normalized gene expression of the subset of genes present in the signature was used to classify the phenotype of interest. For RNAseq input datasets, the normalization consisted in log10 (reads per million mapped read + 1e-7) and genes with less than 20 reads in every sample in the dataset were removed. For microarray input datasets, we retrieved the normalized data from the GEO repository, averaged the normalized signal of all probes per gene and finally used the log10 (average normalized signal per gene + 1e-7) as input for the model. The code used for running the random forest modeling was adapted from https://github.com/jasonzhao0307/R_lib_jason/blob/master/RF_output.R

Given the small sample size of most datasets, the models were trained using leave-one-out cross validation (LOOCV), where for each sample of a dataset, all other samples from the same dataset are used to train the RF model, and the resulting model is used to predict the label or phenotype of the remaining sample. The LOOCV strategy results in one RF model trained per sample per S-D pair. To obtain the combined gene importance feature for a specific S-D pair, the gene importance scores were averaged across all models from a given S-D pair, resulting in one score of "importance" per gene per S-D pair, where the importance measure reflects the mean decrease in node impurity. The receiving operating characteristic (ROC) and precision recall (PR) area under the curve (AUC) are computed using the scores of the single left-out sample per trained model.

Extraction of transfer signatures. Only literature signatures that had a ROC AUC percentile above a given threshold were used at this step. Percentiles were determined as follows: for each S-D pair, 100 random gene lists of the same size were used to compare the performance of the literature signature. Percentiles were used to be able to compare the numbers across datasets that did not have the same case/control distributions. The thresholds of 70, 80 and 90 were empirically tested (**Fig. S9**) and the 70th percentile was chosen, as the two latter were too stringent (in terms of number of literature signatures that passed the threshold) when the signatures were split by group. In order to be able to compare the gene importance feature across literature signatures for a given training dataset, each gene literature signature importance feature was standardized to obtain a mean of 0 and a standard deviation of 1 (z-scores). The z-scores were then aggregated, and the top unique genes were selected as representing the transfer signature.

The number of genes in a transfer signature (N=10, 20 and 50) were empirically tested (**Fig. S3**). The size of 50 genes was chosen for further analyses, with the rationale that (i) 50 genes appeared to provide the best performance in the datasets for which the signature length appeared to play the largest impact and (ii) the larger the signature length the more likely the signature will generalize to other datasets under different conditions. We did however not test transfer signatures containing more than 50 genes for practicality purposes, as the foreseen use of transfer signature will not necessarily be associated with high-throughput sequencing, and a limited size signature has the potential to be more broadly applicable (for example if the markers are assessed through qPCR rather than RNAseq).

Throughout the main text, we used the transfer signature derived from all contributing literature signatures (**Fig. 3-5**), but we also generated and tested transfer signatures based on hallmark and cell type signatures to assess whether they could be broadly applicable, see **Fig S4-9**. The gene lists of transfer signatures are provided in **Dataset S1**.

Gene set overrepresentation was performed on the Biological Process GO ontology. Significance was judged by Benjamini-Hochberg correct p-value cutoff of 0.01. The top 10 significant GO sets are laid out in a plane by placing sets of higher overlap closer to each other. Specifically the 'enrichplot' and 'clusterProfiler' R packages have been used (12). Gene enrichment for Tuberculosis and Dengue transfer signatures are provided in **Dataset S1**.

Prediction in unseen test dataset. Genes identified as markers of "commonality" – present in transfer signatures - were used in an unsupervised analysis to cluster samples from new test datasets, that originated from independent studies (notably new condition, new organism or new infectious agent). The dimension reduction was performed using Uniform Manifold Approximation and Projection (UMAP), followed by Hierarchical Density-Based Spatial Clustering of Application with Noise (HDBSCAN) (13) which can cluster data of varying shape and density. In this approach, the only parameter required is the minimal number of samples per cluster. For this purpose, we tested empirically the minimal number, by identifying the number of samples per cluster that resulted in the lowest number of outliers multiplied by a penalty score equivalent to the square of the number of clusters. This approach limits the creation of excessive numbers of clusters, which could make interpretation difficult. The minimum number of samples per cluster was set to contain at least 7% of the total population. HDBSCAN was run using the *hdbscan* command from the R package "dbscan" (https://github.com/mhahsler/dbscan). The samples considered as outliers by HDBSCAN, were attributed to the closest cluster label using the 3 nearest neighbors with the *knn* command from the R package "dbscan" (https://github.com/mhahsler/dbscan). The code used for running the dimensionality reduction and unsupervised clustering was adapted from https://github.com/NikolayOskolkov/ClusteringHighDimensions/blob/master/easy_scrnaseq_tsne_cluster.R

Once the clusters were identified, the inference of cluster attribution (case or control) was estimated based on the expression of the genes in the signature. Specifically, we did not directly compare the expression between training and test set as the range of expression is most likely more different across dataset than across phenotype within a dataset. We used the direction of the signal rather than the absolute value: for each gene present in the transfer signature, we compared the median expression in each cluster and recorded the direction of the signal in each cluster (high, low or intermediate - in the presence of more than 2 clusters). We performed the same in the training dataset where the transfer signature was obtained from, using the true labels (case/control) instead of clusters to group the samples. We then assessed which cluster in the test dataset had the highest proportion of genes that matched the label of interest in the training dataset (in terms of signal direction) and defined it as “case cluster”, while the other cluster(s) were defined as control cluster. In the rare case where two clusters had the same proportion of matches, we compared the sum of the absolute difference (in median expression) of the genes that matched the direction of the signal in the training dataset. Of note, we need to use biological understanding to decide which phenotype label in the training dataset (**Dataset S1**) would resemble the most the phenotype of interest (“case”) in the test dataset, if not the clusters might be inverted. For example, in the tuberculosis use case, when we used the transfer signatures obtained with the post-challenge timepoint, we expected that the rhesus macaques that were not protected by the vaccine at the end of the study, were the most likely to resemble the individuals that were going to develop acute TB within in a year, as the rhesus macaques were already in a disease state at that time point and the unprotected animals were expected to have a much higher level of immune gene expression in the disease state. While on the opposite, when we used the transfer signatures obtained from the pre-vaccine or pre-challenge datasets, we expected the “case” phenotype to the be rhesus macaques that were protected by the vaccine at the end of the study, as the animals with higher basal level of immune gene expression (such as interferon stimulated genes) are expected to have a higher likelihood of vaccine protection.

References

1. A. Subramanian et al., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 102, 15545-15550 (2005).
2. G. Monaco et al., RNA-Seq Signatures Normalized by mRNA Abundance Allow Absolute Deconvolution of Human Immune Cell Types. Cell Rep 26, 1627-1640 e1627 (2019).
3. S. Bougarn, S. Boughorbel, D. Chaussabel, N. Marr, A curated transcriptome dataset collection to investigate the blood transcriptional response to viral respiratory tract infection and vaccination. F1000Res 8, 284 (2019).
4. 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).
5. J. F. Bermejo-Martin et al., Host adaptive immunity deficiency in severe pandemic influenza. Crit Care 14, R167 (2010).
6. L. M. Franco et al., Integrative genomic analysis of the human immune response to influenza vaccination. Elife 2, e00299 (2013).
7. 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).
8. S. G. Hansen et al., Prevention of tuberculosis in rhesus macaques by a cytomegalovirus-based vaccine. Nat Med 24, 130-143 (2018).
9. M. Liao et al., Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. Nat Med 26, 842-844 (2020).
10. J. Dunning et al., Progression of whole-blood transcriptional signatures from interferon-induced to neutrophil-associated patterns in severe influenza. Nat Immunol 19, 625-635 (2018).

11. D. E. Zak et al., A blood RNA signature for tuberculosis disease risk: a prospective cohort study. Lancet 387, 2312-2322 (2016).
12. G. Yu, L. G. Wang, Y. Han, Q. Y. He, clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS 16, 284-287 (2012).
13. R. J. G. B. Campello, D. Moulavi, J. Sander (2013) Density-Based Clustering Based on Hierarchical Density Estimates. in Pacific-Asia conference on knowledge discovery and data mining, eds J. Pei, V. S. Tseng, C. L., H. Motoda, G. Xu (Springer, Berlin, Heidelberg).

Fig. S1. Study design in detail. This figure complements the study design depicted in **Fig. 1**.Three steps to progress from literature signatures to transfer signatures to prediction in unseen datasets. (**A**) each literature signature (N=148) is used with each training dataset (N=14) as an input to train a random forest model. In other words, there are 148 random forest models per training dataset. (**B**) The gene importance feature and ROC AUC from all random forest models obtained for a given training dataset is used as input to generate one "*transfer signature*" per training dataset. In other words, a single transfer signature is obtained by combining the information obtained from a set of literature gene signatures (here, we start with all literature signatures, excluding the signature coming from the same paper of a given training dataset). (**C**) Finally, the transfer signature derived from each training dataset can be used as an input for unsupervised clustering of an unseen test dataset. UMAP, Uniform Manifold Approximation and Projection. HDBSCAN, Hierarchical Density-Based Spatial Clustering of Application with Noise. LOOCV, leave-one-out cross validation.

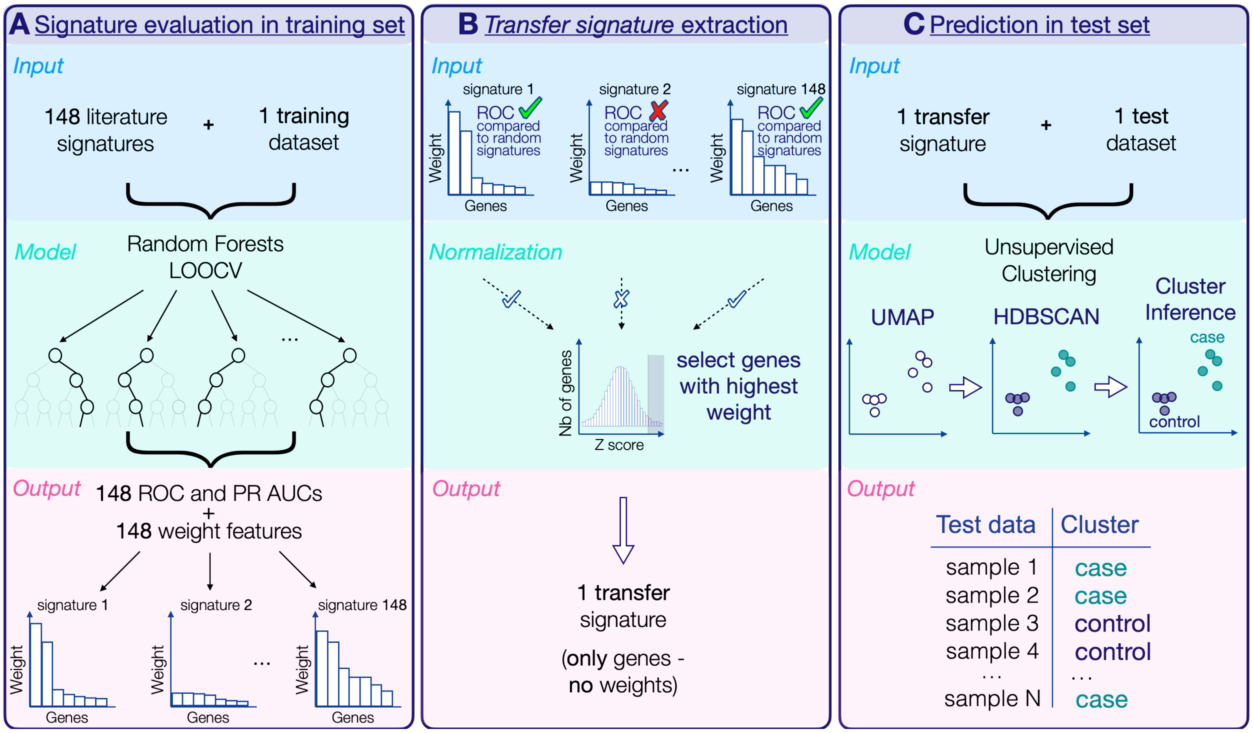


Fig. S2. **Performance of literature signatures as compared to cognate signatures**. The classifying performance of the predicted phenotypes obtained from the random forest leave-one-out cross validation strategies using the literature signatures was assessed for each training dataset (**Methods**, **Dataset S1**). Both panels display the difference in performance (as measured in ROC AUC – **Panel A** – or PR AUC – **Panel B**) between the cognate signature (signature from the same paper than the dataset) and the best performing signature from the literature. When there were multiple signatures originate from the same paper than the training dataset the best performing one was used as "cognate". The literature signatures that outperformed the cognate signature have a positive difference and inversely the ones that did not perform as well have a negative difference. The results are depicted for each group of signatures (**Methods**, **Dataset S1**) – '*global*' encompasses all groups of signatures. The color code is provided in the legend. For both panels, there could be multiple reasons why the cognate signatures do not necessarily perform the best, including (i) the phenotype used for this study may differ from the parent study (f.ex. categorical phenotype rather than numerical; **Dataset S1**), (ii) the cognate signature may only apply to one timepoint of the parent dataset, and finally (iii) some genes from the cognate signature may not have been retrievable due to gene nomenclature conversion issue (**Dataset S1**). In any case, this validates that minor changes in the conditions or analytical settings can alter the results, supporting our strategy to focus on what is shared and transferable across studies rather than what is specific to a single study or condition. ROC, Receiver operating characteristic. AUC, area under the curve. PR, precision recall.

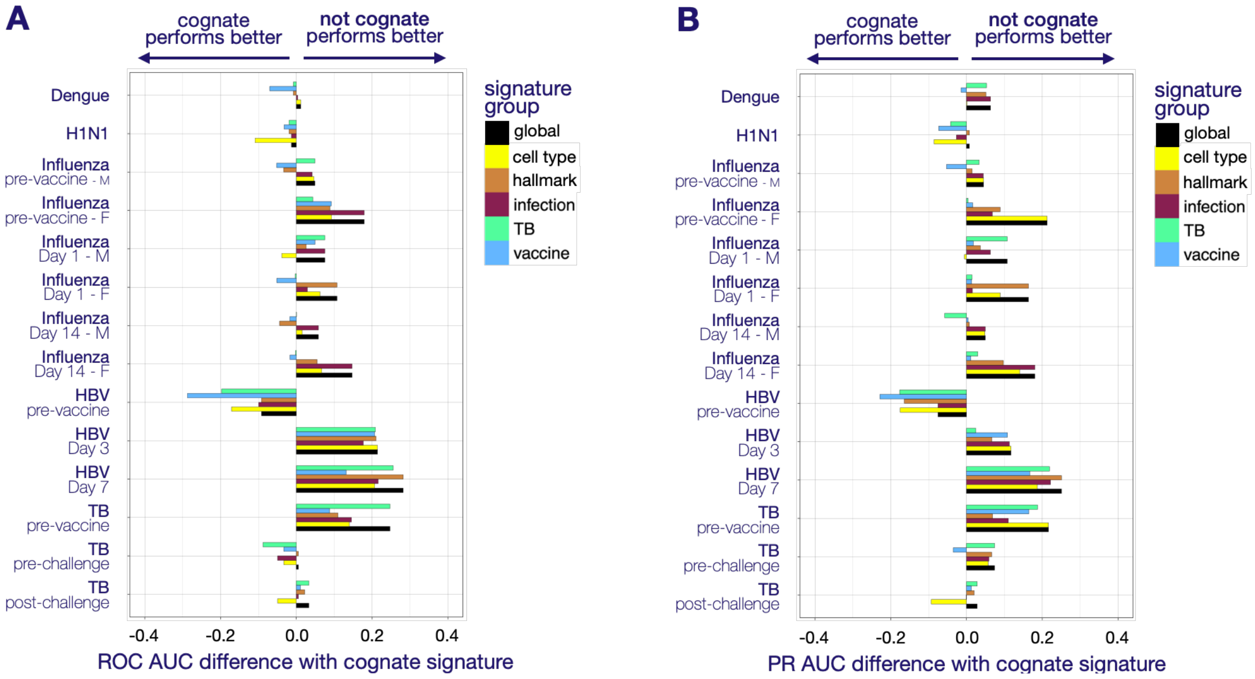


Fig. S3. **Performance of transfer signatures of different sizes**. The classifying performance of the predicted phenotypes obtained from the random forest models (with leave-one-out cross validation) using transfer signatures of varying sizes was assessed for each respective training dataset – where the transfer signatures were obtained from (**Methods**, **Dataset S1**). Three lengths of transfer signatures are depicted in different color and shape. The color code is provided in the legend. **Panel A** displays the ROC AUC obtained for each training dataset. **Panel B** displays the PR AUC obtained for each training dataset. The size of 50 genes was chosen for further analyses, with the rationale that (i) 50 genes appeared to provide the best performance in the datasets for which the transfer signature length appeared to play the largest impact and (ii) the larger the signature length the more likely the signature will generalize to other datasets with different conditions. We did not test transfer signatures containing more than 50 genes for practicality purposes, as the foreseen use of transfer signature will not necessarily be associated with high-throughput sequencing. ROC, Receiver operating characteristic. AUC, area under the curve. PR, precision recall.

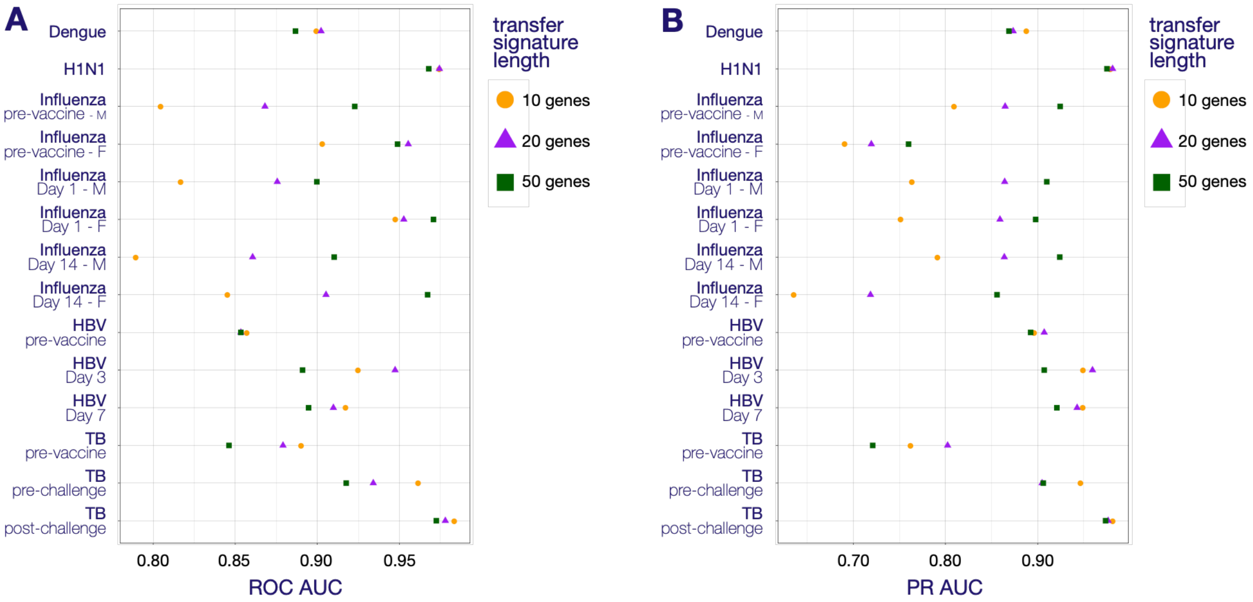


Fig. S4. **Performance of transfer signatures as compared to single signatures**. The classifying performance of the predicted phenotypes obtained from the random forest models (with leave-one-out cross validation) using the transfer or single literature signatures was assessed for each training dataset (**Methods**, **Dataset S1**). Both panels display the difference in performance (as measured in ROC AUC – **Panel A** – or PR AUC – **Panel B**) between the transfer signature and the best single performing literature signature (including the cognate signature for the dataset). The transfer signatures that outperformed the best single literature signature have a positive difference and inversely the ones that did not perform as well have a negative difference. For the purpose of this analysis, we developed not only one transfer signature per training dataset (that was obtained when starting with all literature signatures, **Fig. S1**), but also one transfer signature for the *cell type* and *hallmark* group of signatures, per training dataset. In other words, we started with different subsets of literature signatures to compute the transfer signature and the results are depicted for those three groups of signatures (**Methods**, **Dataset S1**) – '*global*' encompasses all signatures. The color code is provided in the legend. In most instances, the transfer signature outperforms the best performing single signature, with the advantage of increasing the likelihood of generalization in new datasets as transfer signatures are obtained from multiple literature signatures, reducing the risk of extracting condition/study specific markers. ROC, Receiver operating characteristic. AUC, area under the curve. PR, precision recall.

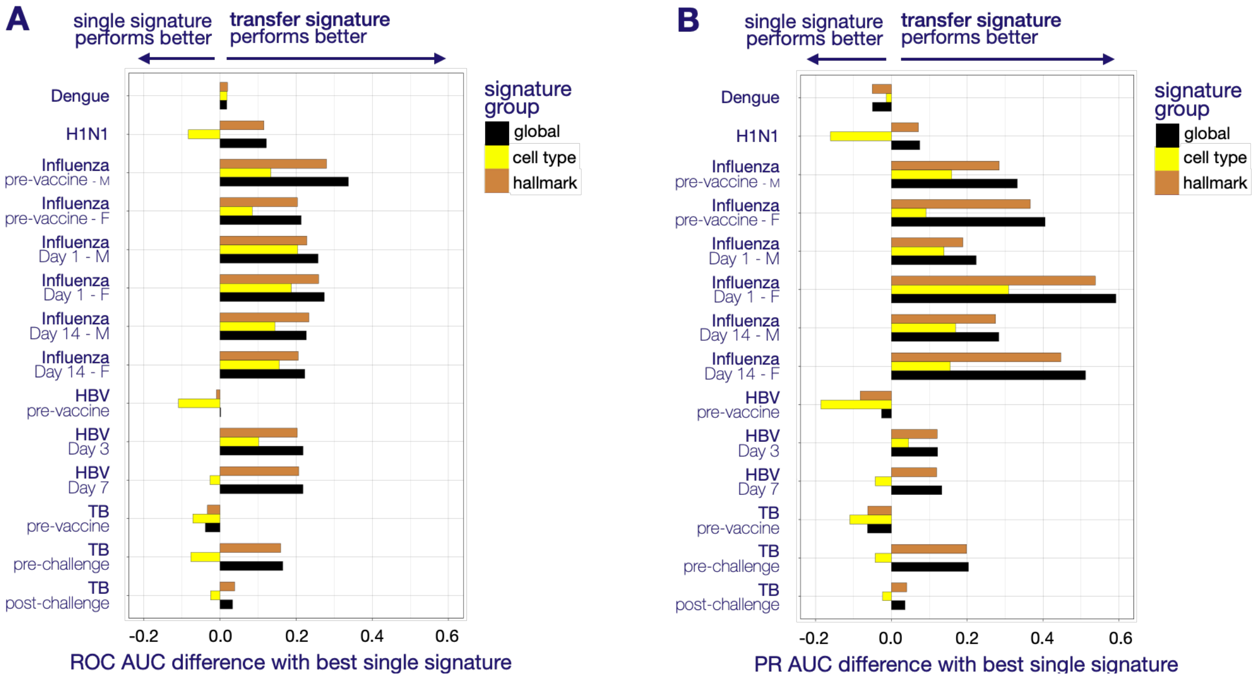


Fig. S5. Tuberculosis progression use case – comparison of transfer signature size performance. **Top panel** shows the study design as displayed in **Fig. 4A**. **Bottom panel** displays the enrichment of cases in the inferred case cluster compared to the other cluster(s) – y axis – using transfer signatures of differing size – x axis. The three plots represent the results obtained with transfer signatures trained with samples obtained at 3 different timepoints shown in the top panel: pre-vaccine, pre-infectious challenge and post-challenge. The results are depicted as boxplot with the individual data overlaid, where each dot represents the result obtained with a transfer signature derived from a different group of literature signatures (global, cell type and hallmark), as explained in **Fig. S4** (see also **Methods**, **Dataset S1**). The enrichment per transfer signature group is further detailed for the 50-gene-long transfer signatures in **Fig. S8**.

Dataset pair description			
Rhesus Macaque - TB vaccine protection study (Hansen et al.)			Training study
Vaccine protection (based on CT score) post challenge; protected (N=13) versus not (N=14)			Training phenotype
Human - latent TB infection longitudinal study (Zak et al.)			Test study/dataset
Developing acute TB within a year (N=30) versus not (N=109)			Test phenotype
TB pre-vaccine	TB pre-challenge	TB post-challenge	Training dataset

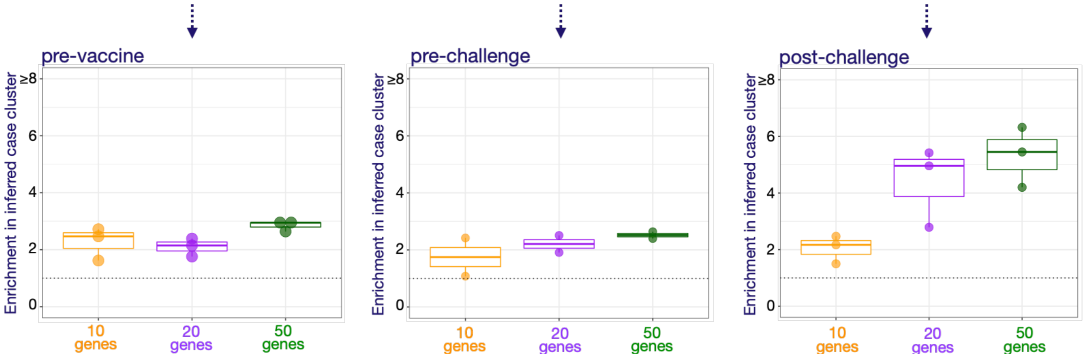
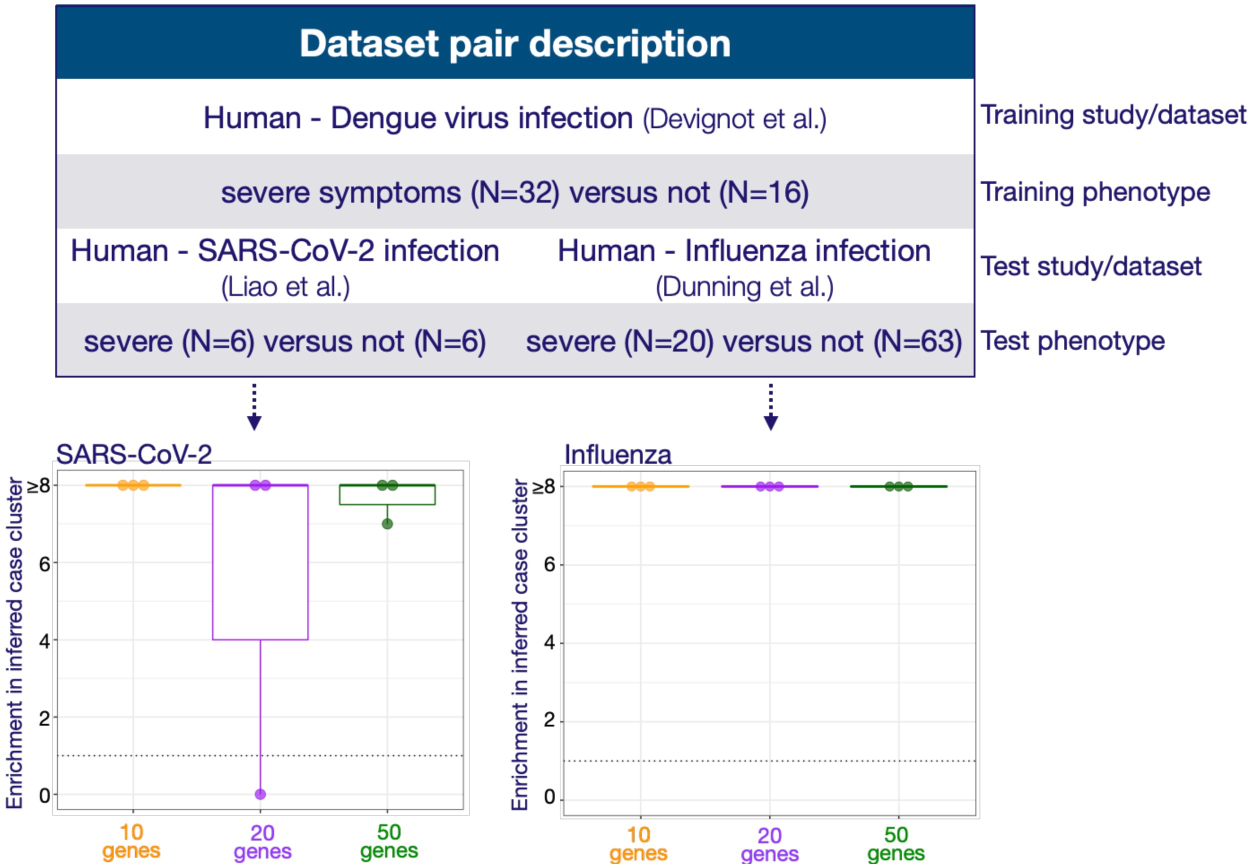
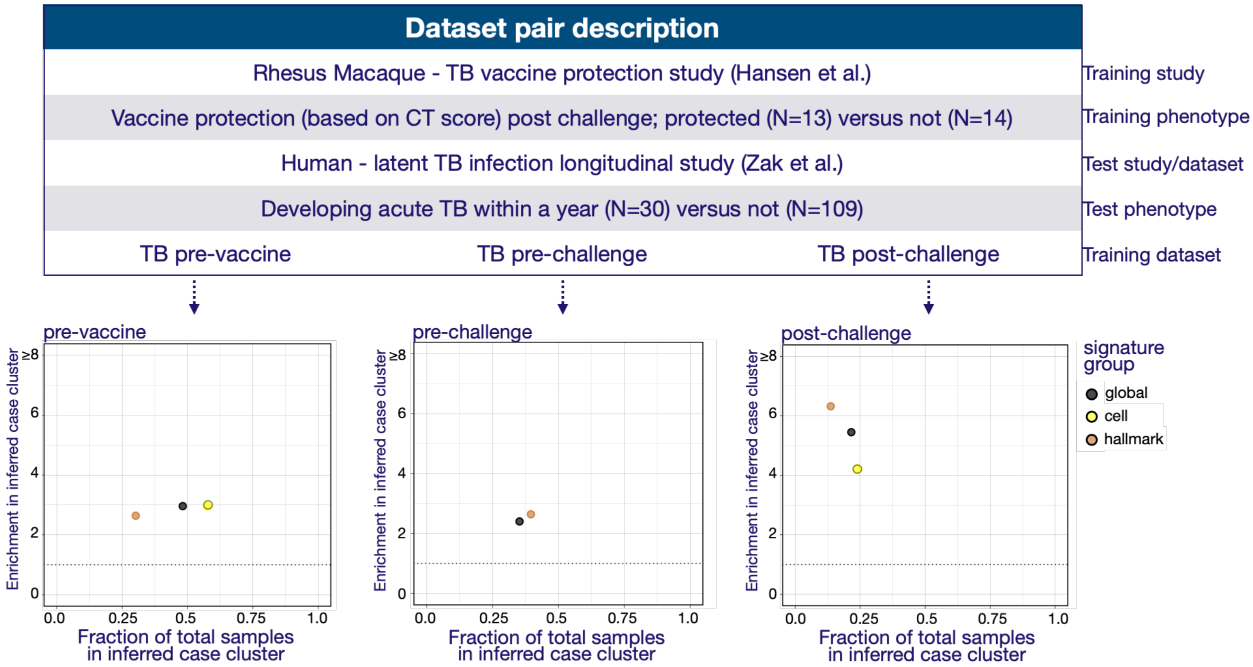


Fig. S6. Severe viral disease use cases – comparison of transfer signature size and performance. **Top panel** shows the study design as displayed in **Fig. 4B**. **Bottom panel** displays the enrichment of cases in the inferred case cluster compared to the other cluster(s) – y axis – using transfer signatures of differing size – x axis. The results are depicted as boxplot with the individual data overlaid, where each dot represents the result obtained with a transfer signature derived from a different group of literature signatures (global, cell type and hallmark), as explained in **Fig. S4** (see also **Methods**, **Dataset S1**). The enrichment per transfer signature group is further detailed for the 50-gene-long transfer signatures in **Fig. S9**. Enrichment below 1 indicates that the "case" cluster was inversely inferred (**Methods**). Here, enrichment depicted as ≥8 indicate that all cases were correctly labeled/present in the inferred case cluster, as seen in **Fig. 4B**.



Dataset pair description		
Human - Dengue virus infection (Devignot et al.)		Training study/dataset
severe symptoms (N=32) versus not (N=16)		Training phenotype
Human - SARS-CoV-2 infection (Liao et al.)	Human - Influenza infection (Dunning et al.)	Test study/dataset
severe (N=6) versus not (N=6)	severe (N=20) versus not (N=63)	Test phenotype

Fig. S7. Tuberculosis progression use case – comparison of transfer signatures obtained from different signature groups. **Top panel** shows the study design as displayed in **Fig. 4A**. **Bottom panel** displays the enrichment of cases in the inferred case cluster compared to the other cluster(s) using 50-gene-long transfer signatures – y axis – versus the fraction of samples present in the inferred case cluster – x axis. The three plots represent the results obtained with transfer signatures trained with samples obtained at 3 different timepoints shown in the top panel: pre-vaccine, pre-infectious challenge and post-challenge. Each dot represents the result obtained with a transfer signature derived from a different group of literature signatures (global, cell type and hallmark) – as explained in **Fig. S4** and where '*global*' encompasses all signatures (see also **Methods**, **Dataset S1**). The color code is provided in the legend. The missing dot for the cell type transfer signature trained on the TB pre-challenge dataset indicates that there were not enough (<50) genes present in the signatures that passed the initial 70th percentile threshold used to extract the transfer signature (**Methods**).



Dataset pair description			
Rhesus Macaque - TB vaccine protection study (Hansen et al.)			Training study
Vaccine protection (based on CT score) post challenge; protected (N=13) versus not (N=14)			Training phenotype
Human - latent TB infection longitudinal study (Zak et al.)			Test study/dataset
Developing acute TB within a year (N=30) versus not (N=109)			Test phenotype
TB pre-vaccine	TB pre-challenge	TB post-challenge	Training dataset

Fig. S8. Severe viral disease use cases – comparison of transfer signatures obtained from different signature groups. **Top panel** displays the study design. **Bottom panel** displays the enrichment of cases in the inferred case cluster compared to the other cluster(s) using 50 gene commonality signatures – y axis – versus the fraction of samples present in the inferred case cluster – x axis. Each dot represents the result obtained with a transfer signature derived from a different group of literature signatures (global, cell type and hallmark) – as explained in **Fig. S4** and where '*global*' encompasses all signatures (see also **Methods**, **Dataset S1**). The color code is provided in the legend. In the SARS-CoV-2 example, due to the small sample size, multiple transfer signatures obtained from different groups of signatures (global and hallmark) generated the same clustering, yielding to the same results in terms of enrichment and fraction and are therefore overlaid and non-visible individually. Here, enrichments depicted as ≥8 indicate that all cases were correctly labeled/present in the inferred case cluster, as seen in **Fig. 4B.**

Dataset pair description		
Human - Dengue virus infection (Devignot et al.)		Training study/dataset
severe symptoms (N=32) versus not (N=16)		Training phenotype
Human - SARS-CoV-2 infection (Liao et al.)	Human - Influenza infection (Dunning et al.)	Test study/dataset
severe (N=6) versus not (N=6)	severe (N=20) versus not (N=63)	Test phenotype

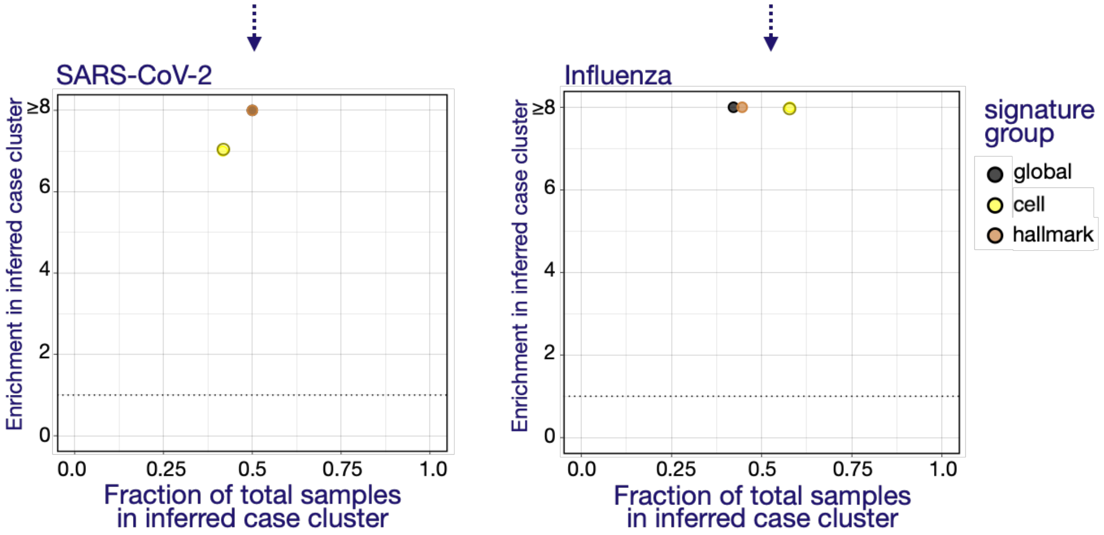
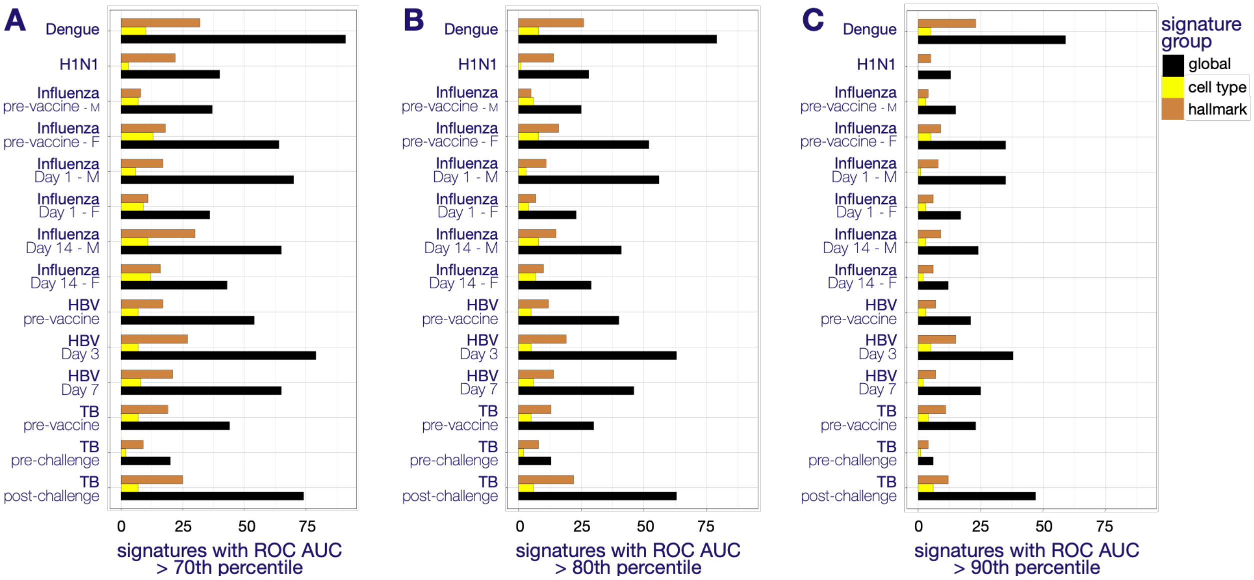


Fig. S9. **Number of literature signatures at different percentile threshold**. The barplots display, for the three groups of signatures used to generate transfer signatures (global, cell type and hallmark), the number of signatures with ROC AUC higher than the 70th percentile (**Panel A**), 80th percentile (**Panel B**) and 90th percentile (**Panel C**) for each signature group. The classifying performance of the predicted phenotypes are obtained from the random forest models (with leave-one-out cross validation) using the literature signatures was assessed for each training dataset. The percentiles are obtained by comparing the literature signature performance to 100 random gene lists of the same size. The higher the percentile, the better the performance of the signature. The color code is provided in the legend. ROC, Receiver operating characteristic. AUC, area under the curve.



Dataset S1. Description of training datasets, testing datasets, literature signatures and transfer signatures. (separate file)

Abbreviation sheet – list of abbreviations used in the different sheets.
Training dataset sheet – description and sources of the training datasets.
Test dataset sheet – description and sources of the test datasets.
Signature description sheet – description and sources of the literature signatures.
Literature ENSG gene lists sheet – ENSEMBL gene ID list for all literature signatures.
Literature ENSG gene list overlap – ENSEMBL gene ID of all genes that appeared in at least one literature signature, as well as the number of signatures they appeared in.
Transfer signature gene lists sheet – ENSEMBL gene ID and gene name list for all transfer signatures
Transfer signature gene list overlap - ENSEMBL gene ID and gene name of all genes that appeared in a least one transfer signature, as well as the number of transfer signature they appear in.
All TS – Enriched Biological Process GO terms for the Dengue and TB transfer signatures.
Dengue TS – Enriched Biological Process GO terms for the Dengue transfer signature for Figure 4.
TB Pre-vaccine TS – Enriched Biological Process GO terms for the TB pre-vaccine transfer signature.
TB Pre-challenge TS – Enriched Biological Process GO terms for the TB pre-challenge transfer signature.
TB Post-challenge TS – Enriched Biological Process GO terms for the TB post-challenge transfer signature for Figure 4.

Dataset S2. Performance of literature signatures compared to random lists of genes. (separate file)

The classifying performance of the predicted phenotypes obtained from the random forest models (with leave-one-out cross validation) using the literature signatures was assessed for each training dataset (**Methods**, **Dataset S1**, **Fig. 1** and **Fig. S1**). The columns represent the training datasets and the rows the literature signatures. In order to be able to compare the performance across datasets (which do not have the same case/control distribution), we evaluated the ROC AUCs in terms of percentiles. The percentiles are obtained by comparing the literature signature performance to 100 random gene lists of the same size. The higher the percentile the better the performance of the signature. Missing data – due to gene conversion issues or no expression in under the curve.

Table S1. Target pairs and non-target pairs of training and test datasets. (separate file)

We define "target pairs" as training-test pairs from diseases with apparent biological relationships. We define "non-target pairs" as training-test pairs from unrelated diseases. All possible pairs of training (n=14) and test datasets (n=3 "target pairs", n=34 "non-target pairs") were evaluated. The table provides the enrichment (**a**) and the F1 score (**b**) obtained when comparing the inferred case cluster versus the inferred control cluster (see step C in **Figure 1**). To facilitate comparison across test datasets with different prevalence of cases and control, for the enrichment metric, the scores are displayed as a rank within each test dataset (rank 1 indicates best performance). The highest score is also provided for each test dataset. The cells that represent an on-target comparison are highlighted in green. The off-target comparison that performed as well as on-target pairs are highlighted in yellow. All dataset descriptions are provided in Dataset **S1**.