Supplementary Material

Materials and Methods

Phosphoflow normalization

We analyzed 89 samples across 13 different plates as previously described (*33*). Fold-change due to stimulation was computed as the ratio of the cell, cytokine stimulation, phospho-protein measure to the raw, un-normalized, cell-phospho-protein matching baseline that was measured on the same plate. Fold-change values were then normalized by the median fold-change difference of a given cell-cytokine stimulation-phospho-protein measure within a given plate. We tested each assay for plate dependent differences and no significant differences between plates were detected post-day normalization.

Gene module construction

As previously described (*33*), of a total of 48,771 gene probes in the microarray per sample, we first selected 6,234 (standard deviation cutoff 0.24) and subsequently normalized their expression by centering and scaling the expression so that each gene's expression across all subjects had euclidean norm equal to 1 for purposes of clustering. We utilized hierarchical agglomerative clustering with average linkage, euclidean distance and height cutoff of 1.5 to derive 109 modules. For each gene module we assigned a set of regulatory genes (regulatory program), based on regression analysis of genes in the modules onto expression of transcription factors using Akaike Information Criterion (AIC) (*70*). To achieve this we used a set of candidate regulators as previously described (*33*). Briefly, we performed linear regression with elastic net penalty of each module's expression onto the set of regulators using LARS-EN algorithm with *l*2 penalty weighted by 0.01. The LARS-EN algorithm provides fits of increasing number of predictors. In order to select the best model among the outputs of LARS-EN we assessed quality of the resulting models by AIC, with sample specific terms weighted by within module variance. The fit with the best AIC score was selected for each module. Gene modules and their regulatory programs can be accessed at http://www.cs.unc.edu/~vjojic/fluy2-upd/.

Cross-validation and feature selection for finding immune-signatures of latent CMV and aging

I. Power analysis

Given sample size and 6 related tasks of sizes corresponding to classification tasks between 4 age/cmv combinations, an effect size exceeding 0.5 is detected with probability 73.03%. We synthesized data with randomly distributed weights across 6 tasks such that rank of weight matrix is 2. We generated 100 such synthetic datasets. We run our method 100 times on these datasets. We computed power of our method to detect an effect in excess of 0.5.

II. Regularization procedure

An integral part of our training algorithm is a procedure for fitting multiple logistic regression models with elastic net (*71*) and nuclear norm (*72*) penalties. In our case there are 6 related tasks that arise from contrasting pairs of populations yCMV, yCMV+, oCMV-, oCMV+. The *l*1 factor from elastic net penalty encourages discovery of concise predictors in each of the tasks separately. The nuclear norm penalty promotes reuse of the same predictors across tasks. This is accomplished by lowering the rank of the matrix of all predictor weights across all tasks. The optimization cost can be stated as:

$$\sum_{c=1}^{C}\left[\frac{1}{n_c}\left(\sum_{t=1}^{n_c}\log\left(1+\exp\left(-y_{c,t}\left(\beta_c^T x_{c,t}+\alpha_c\right)\right)\right)\right)+\lambda\sum_{i=1}^{p}\left|\beta_{c,i}\right|+\gamma\sum_{i=1}^{p}\beta_{c,i}^2\right]$$
$$+\mu\|[\beta_1 \dots \beta_C]\|$$

where C is the number of tasks, p is the number of predictors, n_{c} is the number of samples for task c, $x_{c,t}$ is the vector of predictor values for subject t in task C, and $y_{c,t}$ indicates one of the two populations relevant to the task c. In addition to the logistic regression loss, the objective above contains three penalties. These penalties are l1, a sum of absolute values of predictor weights, ridge, a sum of squares of predictor weights, and nuclear norm, a sum of singular values of matrix constructed by stacking predictor weight vectors β for the C tasks. Each of these penalties is weighted by a parameter, λ, γ and μ above. This objective is optimized using an Alternating Direction Method of Multipliers. We assume all of our predictors are standardized to mean 0 and standard deviation 1. The result of our fitting procedure is a matrix of predictor weights β of size p x C, with one column per task, and vector of task specific intercepts α for the logistic regression model. In practice, penalty weights λ, γ and μ in Eq. 1 are set by a data driven procedure, such as cross-validation. We applied 3-fold cross-validation with the parameters chosen to yield the lowest average error across the tasks while using the smallest set of predictors. We did not test for normality or for heteroscedasticity since logistic regression does not make assumption of normality nor imply equal variance across groups. The computer source code can be found at https://github.com/vjojic/CMVAge.

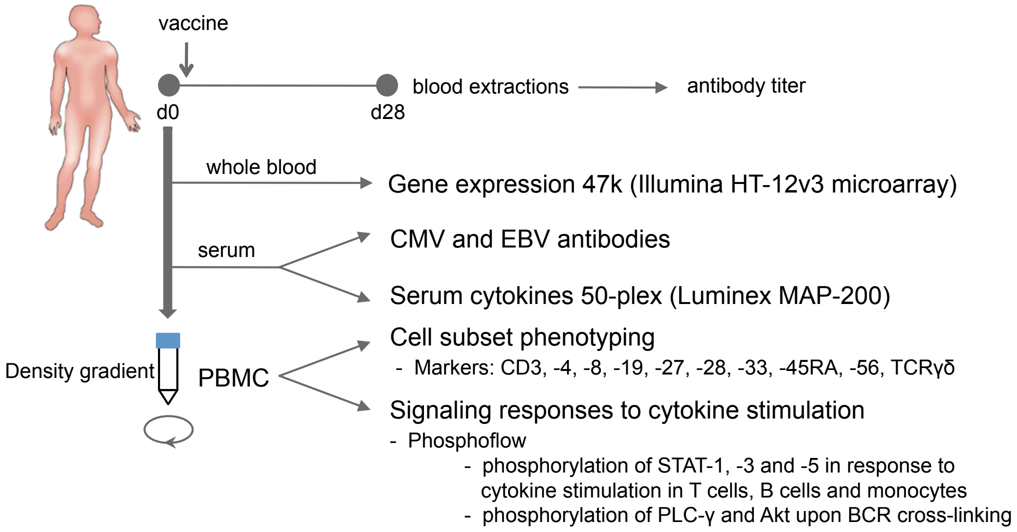
Association of SNPs and cell subset frequency

We first narrowed our analysis to SNPs with minor allele frequency, in our dataset, greater than 5% (116,405 SNPs). On this set we performed fisher exact test of association between SNP allele and the frequency of $CD4^{+}$ $CD28^{-}$ cells controlling for CMV seropositivity. We performed FDR correction of the resulting p-values.

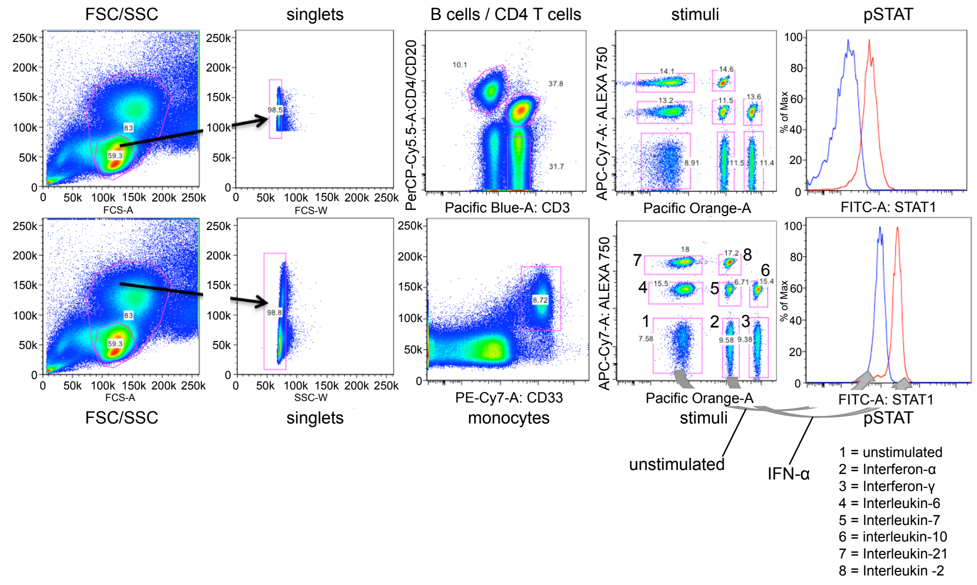
Gene enrichment analysis

Genes from each module were explored by using Ingenuity Pathway Analysis (IPA) for function enrichments. Data was imported and Core Analysis was performed with the following setting: Data Source: Ingenuity Expert Findings; Confidence: Experimentally Observed, TarBase, Protein-protein Interactions, Additional interactions; Species: Human. Most significant function enrichment for each module were explored in selected modules and used for quality control. Functional enrichments with highest significance are reported. In addition, all genes in these modules and module regulators with highest regression coefficient or with known function were manually curated using a variety of sources including PubMed, IPA and BIOBASE Knowledge Library.

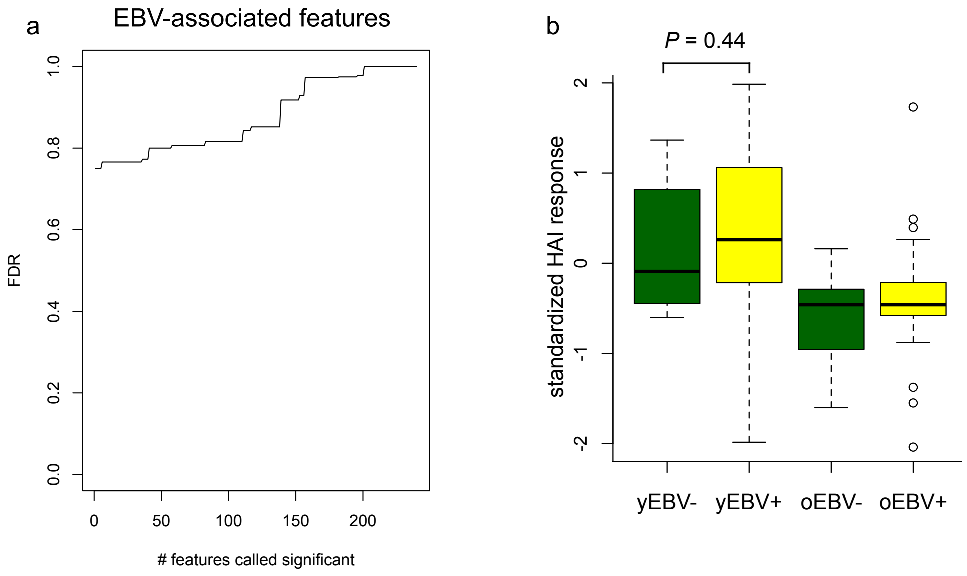
Supplementary Figures and Tables


vaccine
d0
d28
blood extractions
antibody titer
whole blood
Gene expression 47k (Illumina HT-12v3 microarray)
serum
CMV and EBV antibodies
Serum cytokines 50-plex (Luminex MAP-200)
Density gradient
PBMC
Cell subset phenotyping
- Markers: CD3, -4, -8, -19, -27, -28, -33, -45RA, -56, TCRγδ
Signaling responses to cytokine stimulation
- Phosphoflow
- phosphorylation of STAT-1, -3 and -5 in response to cytokine stimulation in T cells, B cells and monocytes
- phosphorylation of PLC-γ and Akt upon BCR cross-linking

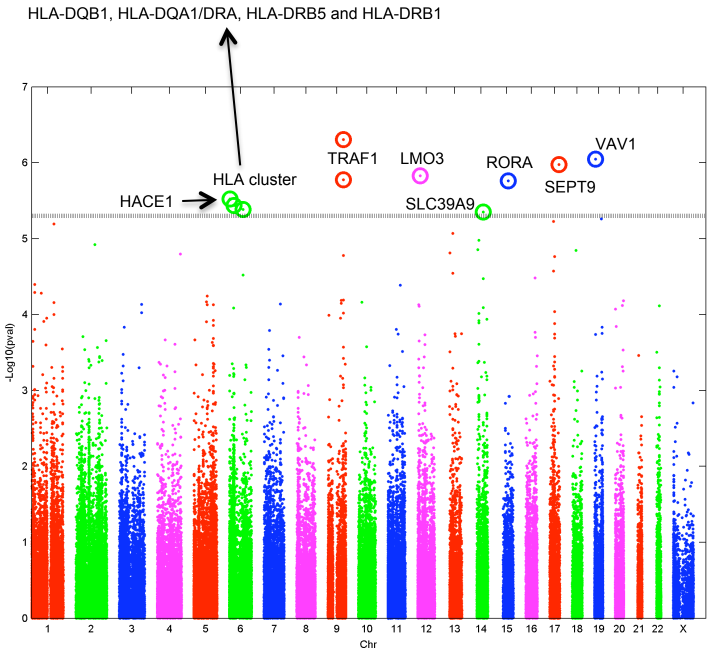
Supplementary Figure 1. Study design. Blood samples are obtained before (d0) and 28±7 days (d28) after a single intramuscular inoculation of the seasonal inactivated influenza vaccine. Samples from d0 are used for gene expression analysis, determination of serum cytokines and chemokines, and presence of CMV/EBV antibodies, cell subset phenotyping and signaling responses to cytokine stimulations on CD4(+) and CD8(+) T cells, B cells and monocytes as well as the phosphorylation of PLC-γ and Akt upon BCR cross-linking on B cells. Serum samples from d0 and d28 are utilized for determination of anti-influenza antibody titers by the hemagglutinin inhibition assay.



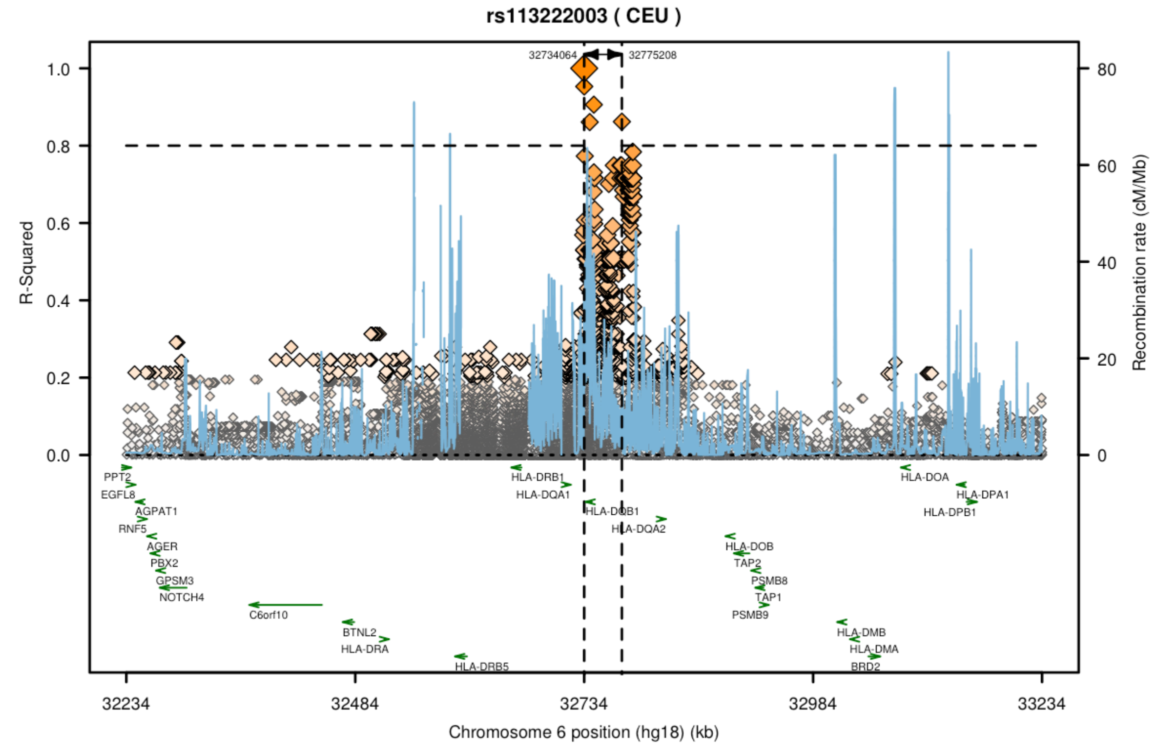
Supplementary Figure 2. Gating strategy for phosphoflow assays. Phosphorylation of STAT proteins before and upon stimulation with cytokines is analyzed using FlowJo software by gating on live cells as discriminated by FSC/SSC profiles, then using double gating for singlet discrimination, followed cell subset-specific gating. Phosphorylation of STAT1, 3, and 5 proteins is analyzed by deconvolution of stimuli-specific gating.



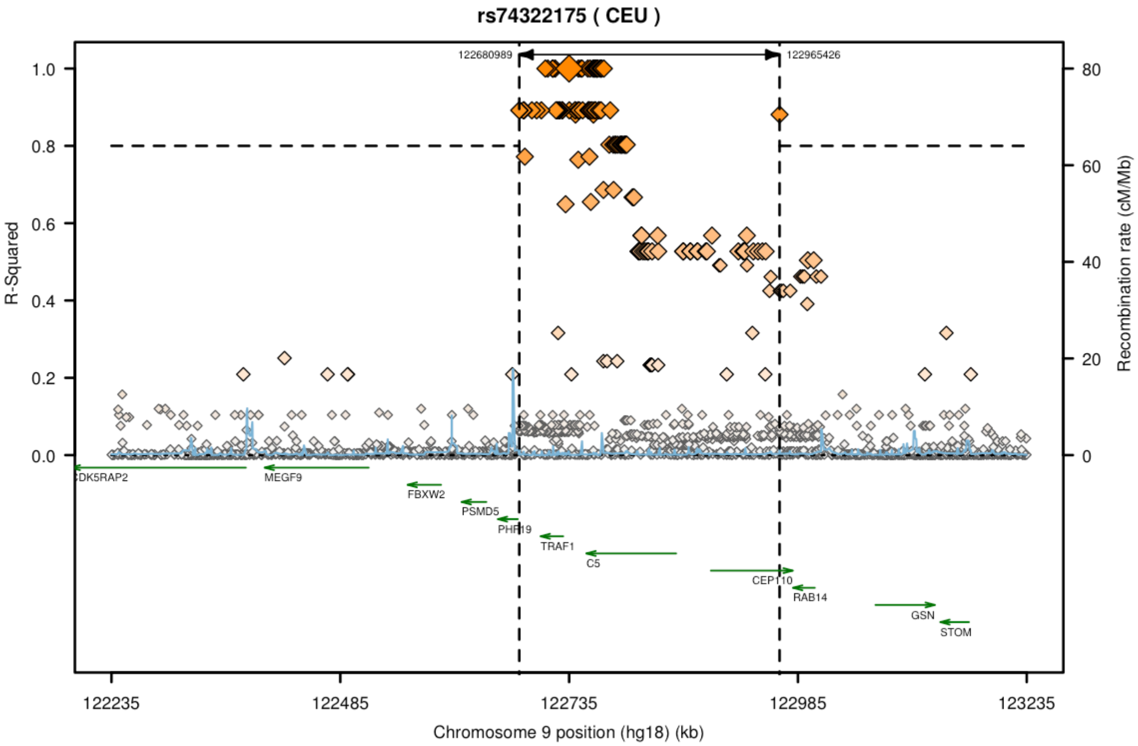
Supplementary Figure 3. No significant effect of EBV in immune measures and response to influenza vaccine. Multiple regression analysis for a total of 236 features against EBV seropositivity, age and sex was performed and significance obtained via permutation tests (*73*). As depicted in the figure the feature with the highest significance for EBV is detected at a FDR *Q* = 0.75 (a). Vaccine responses were measured as the delta geometric mean titer for all three strains and standardized for visualization purposes. No differences are observed between yEBV- and yEBV+ (b).



Supplementary Figure 4. Manhattan plot showing genetic variants that associate with CMV-related phenotypic alteration. The frequency of CD4^{+} CD28^{-} cells, a hallmark of CMV infection, was correlated against a restricted set of genetic polymorphisms (SNPs) previously identified in several immune-related pathologies (*47*). 35 SNPs were significantly associated with the frequency of CD4^{+} CD28^{-} cells at an FDR < 5% ($P < 5 \times 10^{-6}$, FDR $Q < 0.05$) (dotted line). Notable SNP candidates include those in the vicinity of HLA genes, *TRAF1*, *RORA, C5* and *VAV1*.



Supplementary Figure 5. SNAP plot of notable SNPs found to be associated with the CD4^{+} CD28^{-} cell frequency on chromosome 6. On a total of 116,405 SNPs, fisher exact test of association between SNP allele and the frequency of CD4^{+} CD28^{-} cells controlling for CMV seropositivity was performed. Figure shows important SNPs on chromosome 6 close to HLA genes. SNAP plot was generated using 1000 genomes Pilot 1, Panel = CEU, r^{2} threshold = 0.8.



Supplementary Figure 6. SNAP plot of notable SNPs found to be associated with the CD4^{+} CD28^{-} cell frequency on chromosome 9. On a total of 116,405 SNPs, fisher exact test of association between SNP allele and the frequency of CD4^{+} CD28^{-} cells controlling for CMV seropositivity was performed. A cluster of 27 SNPs spanning a region in chromosome 9 from position 122,705,118 to 122,748,094 is located in the vicinity and within *TRAF1* and 55kb centromeric from *C5* is shown. SNAP plot was generated using 1000 genomes Pilot 1, Panel = CEU, r^{2} threshold = 0.8.

Baseline characteristic	*Young Y1*	*Older Y1*	*Young Y2*	*Older Y2*	*Cohort 2*
Age range (median)	20-30 (24.5)	61->89 (78)	22-32 (26)	62->89 (77)	19-44 (27)
Gender					
Male	16 (53%)	21 (34%)	16 (64%)	18 (35%)	10 (27%)
Female	14 (47%)	40 (66%)	9 (47%)	34 (65%)	27 (73%)
Cytomegalovirus (+)	57%	59%	55%	60%	51%
Epstein Barr Virus (+)	53%	67%	57%	65%	N/A

Supplementary Table 1. Subjects' baseline characteristics

	Immune variable	yCMV- vs yCMV+	yCMV- vs oCMV-	yCMV- vs oCMV+	yCMV+ vs oCMV-	yCMV+ vs oCMV+	oCMV- vs oCMV+
CYTOKINES/CHEMOKINES	EOTAXIN	0	0.42	0.059	0.27	0.048	-0.032
	IFN-g	0.093	0	0	0	0	0
	IL-12P40	0	0	0	0	0.0084	0
	IL-13	0.06	0	0	0	0	0
	IL-1ra	0	0	0	0	0	0.08
	IL-5	0	0.039	0	0	0	0
	IP10	0	0.25	0.2	0.016	0.12	0
	MCP-1	0	0.022	0	0	0	0
	TGF-b	0	0	0	0	0	-0.02
CELL SUBSETS	CD4	0	0	0	0.012	0	0
	CD4.TEM	0.31	0	0	0	0	0
	CD4.Naive	0	0	-0.059	0	0	0
	CD4.EM	0.1	0	0	0	0	0
	CD8	0	-0.45	0	-0.58	-0.014	0
	CD8.TEM	0.62	0.17	0.067	-0.075	0	0.39
	CD8.Naive	0	-0.26	-0.62	0	-0.32	-0.23
	CD8.CM	-0.082	0	0	0.21	0.076	0
	NKT.cells	0.38	0	0	0	0	0
	CD4.CD28-	0.047	0	0	-0.21	0	0.08
	CD8.CD28-	0.072	0.13	0.19	0	0	0
SIGNALING NODES	cd20.base.STAT5	-0.0013	0	0	0	0	0
	cd8.base.STAT1	0	0	0	0.13	0.17	0
	cd20.IFNa.STAT1	0	0	0	0	0	0.092
	cd20.IFNg.STAT1	0	0	0	0	0	0.027
	cd20.IL6.STAT1	0	-0.0038	0	0	0	0
	cd4.IFNg.STAT3	0.025	0	0	0	0	0
	cd4.IL6.STAT5	0.089	0	0	-0.16	-0.21	0
	cd8.IFNa.STAT1	0	0	0	0	-0.019	0
	cd8.IFNa.STAT3	0	0	0	0	0	-0.062
	cd8.IFNa.STAT5	0	0	0	0	0	-0.014
	cd8.IFNg.STAT1	0	-0.053	-0.12	-0.22	-0.24	0
	cd8.IL6.STAT1	0.072	0	0	-0.19	-0.26	0
	cd8.IL6.STAT3	0.23	0	0	0	-0.093	0
	cd8.IL6.STAT5	0	-0.023	-0.19	-0.063	-0.13	0
	cd8.IL10.STAT3	0	0.021	0	0	0	0
	cd8.IL10.STAT5	0	0	0	0.088	0	-0.11
	cd8.IL21.STAT1	0	-0.077	-0.038	0	0	0
	mono.IFNg.STAT3	0	0	0	0	-0.14	-0.058
GENE MODULES	mod_007	0	0	0	0	-0.016	0
	mod_020	0	0	0	0.085	0	0
	mod_029	0	0	0	0	0	-0.068
	mod_030	0.021	0	0	0	0	0
	mod_034	0	-0.017	-0.051	0	0	0
	mod_035	0	-0.021	0	0	0	0
	mod_038	0	0	0	-0.0055	0	0.26
	mod_039	0	-0.12	-0.078	0	0	0
	mod_041	0	-0.0018	-0.21	0	-0.097	0
	mod_043	0	-0.022	-0.013	0	0	0
	mod_047	0	-0.024	0	0	-0.0084	0
	mod_054	0	-0.066	0	0	0	0
	mod_077	0	0	0	0	0	0.15
	mod_098	0	-0.13	-0.0022	0	0	0
	mod_101	0	-0.13	0	0	0	0
	mod_102	0	-0.05	0	0	0	0
	mod_103	0.2	0	0	0	0	0
	mod_108	0	-0.14	-0.31	-0.18	-0.34	0
	% accuracy (baseline)	79 (55)	91.7 (65)	95.7 (73)	88 (60)	90.2 (69)	63 (60)

Supplementary Table 2. Immune parameters computationally selected in all six classification problems. Red = positive regression coefficient, Blue = negative regression coefficient. yCMV- = young cytomegalovirus (-), yCMV+ = young cytomegalovirus (+), oCMV- = old cytomegalovirus (-), oCMV+ = old cytomegalovirus (+).