Early post-vaccination blood transcriptomics predicts adaptive immune outcomes to the yellow-fever vaccine YF-17D, but the predictive genes dissociate from canonical antiviral signatures: a leakage-controlled machine-learning re-analysis

K-Dense Web

K-Dense Web Research

contact@k-dense.ai

July 10, 2026

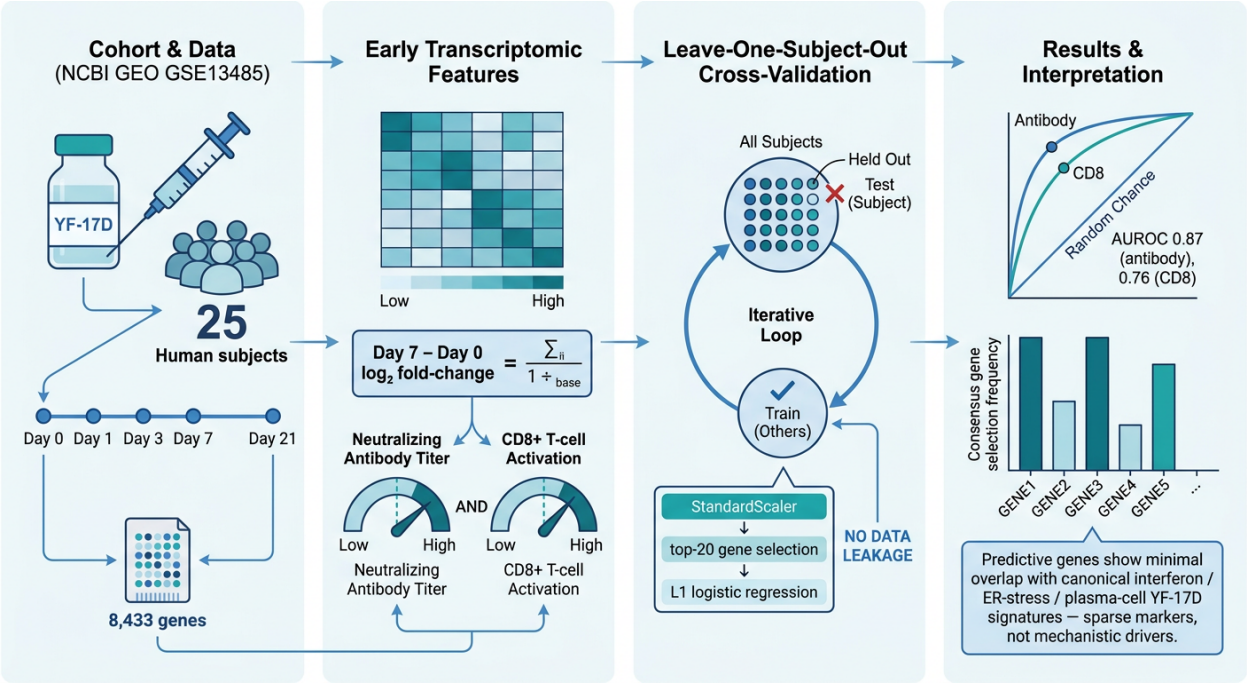

Figure 1: Graphical abstract. Blood-transcriptome time courses from 25 healthy adults vaccinated with the live-attenuated yellow-fever vaccine YF-17D (NCBI GEO GSE13485) were used to predict two dichotomized adaptive-immune endpoints—neutralizing-antibody titer and $CD8^+$ T-cell activation—from early ($\leq$ Day 7) gene expression. Features were the Day 7 − Day 0 $\log_2$ fold-change (LFC) or raw Day 7 expression over 8,433 genes. A strictly nested Leave-One-Subject-Out Cross-Validation (LOOCV) pipeline (fold-internal scaling, univariate filtering, and L1-penalized logistic regression) achieved LOOCV AUROC of 0.87 (antibody) and 0.76 ($CD8^+$). The consensus predictive genes showed minimal overlap with the canonical interferon/ER-stress/plasma-cell YF-17D signatures, and are therefore framed as sparse predictive markers rather than validated mechanistic drivers.

Abstract

Background. Systems vaccinology posits that the early innate transcriptional response to vaccination encodes information about the eventual adaptive immune outcome. The landmark study of Querec and colleagues (2009) showed that early blood-transcriptome signatures could predict immunogenicity of the live-attenuated yellow-fever vaccine YF-17D. Re-analysing this canonical cohort with a modern, leakage-controlled machine-learning workflow provides an opportunity to test both the predictive claim and the interpretability of the features that drive prediction.
Methods. We retrieved the YF-17D human vaccination time course (NCBI GEO GSE13485; $N = 25$ subjects across two trials; Days 0, 1, 3, 7, 21) together with per-subject neutralizing-antibody titers and $CD8^+$ T-cell activation measurements. After probe-to-gene collapse (8,433 unique genes), we defined two early-timepoint feature sets—the within-subject Day 7 − Day 0 $\log_2$ fold-change (LFC) and raw Day 7 expression—and dichotomized each endpoint at its median. Predictive models were evaluated with Leave-One-Subject-Out Cross-Validation (LOOCV, 25 folds) in which standardization, univariate feature filtering (top-20 genes by point-biserial correlation), and an L1-regularized logistic classifier were fit strictly inside each training fold. We quantified feature-selection stability (fold-wise Jaccard index, consensus genes) and compared consensus genes to curated YF-17D signatures using over-representation analysis (Enrichr and a background-matched hypergeometric test).
Results. Both endpoints were predicted above the no-information rate (0.52). Neutralizing-antibody class was best predicted by baseline-corrected LFC features (LOOCV AUROC = 0.87, accuracy = 0.80, sensitivity = 0.77, specificity = 0.83); $CD8^+$ T-cell activation was best predicted by raw Day 7 expression (AUROC = 0.76, accuracy = 0.72). Feature selection was moderately stable (mean pairwise Jaccard $\approx$ 0.48–0.55), yielding 10–13 consensus genes selected in $\geq$80% of folds per model. Strikingly, of 21 high-confidence consensus genes, only *IL5* mapped to a canonical YF-17D immune module; none overlapped the interferon-stimulated-gene, ER-stress/unfolded-protein-response, or plasma-cell (*TNFRSF17*) signatures that dominate the published YF-17D response.
Conclusions. Early post-vaccination transcriptomics carries a genuine, cross-validated signal for YF-17D adaptive-immune outcomes, but the sparse features selected in this high-dimensional (P = 8,433), low-sample-size ($N = 25$) setting are predictive markers rather than validated mechanistic drivers, and largely bypass the canonical antiviral biology. We report per-subject predictions, feature-stability diagnostics, and fully reproducible code, and argue for independent-cohort validation and module-level analysis before any biological claim.

Keywords: systems vaccinology; yellow fever; YF-17D; blood transcriptomics; machine learning; cross-validation; feature-selection stability; overfitting.

1 Introduction

Vaccination is among the most successful interventions in the history of medicine, yet the mechanistic determinants of vaccine-induced immunity remain incompletely understood, and the magnitude of the adaptive immune response varies substantially between individuals [1, 2]. A central aspiration of the field of *systems vaccinology* is to identify early, easily measurable molecular correlates—particularly the transcriptional program mobilized in peripheral blood within the first days after immunization—that predict, and ideally explain, the eventual adaptive immune outcome [1, 2]. If such early signatures exist and generalize, they would enable rapid immunogenicity read-outs, accelerate iterative vaccine design, and illuminate the innate-to-adaptive circuitry that governs durable protection.

The live-attenuated yellow-fever vaccine YF-17D is the archetypal model system for this program. Developed in the 1930s and administered to hundreds of millions of people, YF-17D is one of the most effective vaccines ever produced: a single dose elicits broad, polyfunctional, and long-lived

neutralizing-antibody and $CD8^+$ T-cell responses [3, 4]. Mechanistic work established that YF-17D activates multiple dendritic-cell subsets through a combinatorial engagement of Toll-like receptors (TLR2, TLR7, TLR8, and TLR9), thereby orchestrating a balanced T-helper-1/T-helper-2 response and priming a remarkably diverse effector repertoire [5, 6]. In humans, YF-17D triggers an integrated, multilineage innate response detectable in blood, characterized prominently by a type-I interferon and antiviral program, complement activation, and, by the first post-vaccination week, signatures of proliferating effector lymphocytes [7]. The magnitude of the vaccine-induced $CD8^+$ T-cell response is itself substantial and polyfunctional, peaking around the second post-vaccination week [4].

Against this backdrop, the landmark study of Querec et al. [8] applied a systems-biology approach to a cohort of healthy adults vaccinated with YF-17D, profiling the blood transcriptome longitudinally alongside measured adaptive-immune endpoints. That study reported early gene-expression signatures—including a module linked to the integrated stress response/eIF2α signaling (e.g. *EIF2AK4*) and an innate-sensing/complement module—that could discriminate high from low responders for both neutralizing antibody and $CD8^+$ T-cell responses, and it validated candidate predictors in an independent trial [8]. This work, together with the contemporaneous characterization of the YF-17D response by Gaucher et al. [7], catalysed the modern systems-vaccinology literature: analogous early-signature predictors were subsequently reported for seasonal influenza vaccination [9] and, in a cross-vaccine meta-analysis that also introduced the widely used Blood Transcription Modules (BTMs), for five distinct human vaccines [10].

The claim that early transcriptomics predicts adaptive immunity is now foundational, but it rests on analyses conducted with the sample sizes that human vaccine trials realistically permit—typically a few dozen subjects assayed over tens of thousands of transcripts. This is the classic "large-P, small-N" regime, in which naive supervised learning is notoriously prone to over-optimistic performance estimates and to selecting unstable, non-reproducible feature sets [11, 12]. Two distinct methodological hazards are relevant. First, if any step that uses the outcome—feature selection, standardization to outcome-informed statistics, or hyperparameter tuning—is performed on the full dataset before cross-validation, information leaks from the test folds into training, inflating apparent accuracy [11, 12]. Second, even with a scrupulously nested evaluation, the particular genes chosen by a sparse classifier can vary substantially across resampling folds, so that a model may predict well while its "signature" is statistically arbitrary [13, 14]. Distinguishing a genuine predictive signal from an interpretable mechanistic signature therefore requires both leakage-free performance estimation *and* an explicit assessment of feature-selection stability.

Here we revisit the canonical YF-17D cohort with these hazards squarely in view. Working from the publicly deposited data (NCBI GEO GSE13485), we (i) locate the transcriptome time course and the two adaptive-immune endpoints; (ii) define a principled early time point (Day 7) and two response endpoints (neutralizing-antibody titer and $CD8^+$ T-cell activation), each dichotomized at its median; (iii) build predictive models using a strictly nested Leave-One-Subject-Out Cross-Validation (LOOCV) pipeline that prevents data leakage; (iv) report discrimination (AUROC) and accuracy relative to the no-information rate, together with an explicit stability analysis of the selected features; and (v) enumerate the genes and pathways the models rely on and compare them, honestly, to the antiviral/innate signatures established for YF-17D [7, 8]. Our central finding is a deliberate juxtaposition: early transcriptomics does carry a real, cross-validated predictive signal, yet the sparse features selected under this design largely *dissociate* from the canonical YF-17D biology—a result we interpret not as a contradiction of the systems-vaccinology paradigm but as a concrete illustration of the difference between a predictive marker and a mechanistic driver in the high-dimensional, low-N setting.

2 Methods

All analyses were implemented in Python 3.12 using `pandas`, `numpy`, `scipy`, `statsmodels`, and `scikit-learn` [15]; every stochastic step used a fixed random seed (42), and all code, intermediate tables, and figures are provided (see *Code and Data Availability*).

2.1 Data acquisition and curation

The YF-17D human vaccination time course was retrieved programmatically from the NCBI Gene Expression Omnibus [16, 17] as series matrix files for GSE13485 and its SuperSeries companion GSE13486. Sample characteristics (time point, trial, neutralizing-antibody titer, T-cell activation, tissue) were parsed into typed columns, and subject identifiers and post-vaccination day were derived from sample titles. The in-vivo dataset comprises 87 microarray samples from **25 subjects** distributed across two trials (Trial 1: 15 subjects, Days 0/1/3/7/21; Trial 2: 10 subjects, Days 0/3/7) on the custom Agilent-based platform GPL7567 (Table 1; Fig. 2). Two adaptive-immune outcomes are recorded per subject: the neutralizing-antibody titer and a $CD8^+$ T-cell activation index.

Data-integrity safeguard. A verification of GSM accessions revealed that the GPL7567 component of GSE13486 re-hosts the *identical* 87 arrays as GSE13485 (100% accession overlap); it is therefore a duplicate SuperSeries record and must not be treated as an independent validation cohort. Accordingly, the genuine discovery→validation partition is *Trial 1 versus Trial 2 within GSE13485*, and all cross-trial analyses use this axis. The only unique data contributed by GSE13486 is a small in-vitro YF-17D stimulation subset (GPL7566; 10 arrays, 2 donors) that is not used for outcome modeling.

2.2 Preprocessing and probe-to-gene mapping

The deposited expression values spanned [2.78, 15.15] and were thus already $\log_2$-transformed; per-sample medians were tightly aligned (6.089 ± 0.041; IQR 3.07 ± 0.14) across all 87 arrays and both trials, so the authors' normalized intensities were used directly as the primary matrix (quantile normalization was computed as a documented alternative but not applied, to avoid re-normalizing already-normalized data). Probes were annotated to HGNC symbols via GenBank accessions using the mygene.info service; of the 8,537 probes with a resolved symbol, genes with multiple probes were collapsed by retaining the highest-mean-expression probe (the deterministic maxMean rule). This yielded a gene-level matrix of **8,433 unique genes** $\times$ 87 samples. A principal-component analysis showed a mild Trial 1 versus Trial 2 separation along PC1 (23.5% variance; Fig. 3), motivating a baseline-corrected feature representation and strict subject/trial-level partitioning downstream.

2.3 Feature construction and endpoint definition

All 25 subjects had paired Day 0 and Day 7 samples. **Day 7** was selected as the early post-vaccination time point because it is the first post-vaccination day sampled in *both* trials at full cohort size and lies at the transition between the innate transcriptional burst and the emergence of proliferating effector lymphocytes described for YF-17D [7, 8], while preceding the antibody/$CD8^+$ effector peak. Two complementary early-timepoint feature sets, each spanning the 8,433 genes, were constructed:

- **LFC**: the within-subject Day 7 − Day 0 $\log_2$ fold-change (a log-ratio that removes each subject's baseline offset and mitigates the trial batch effect); and
- **Day7**: the raw Day 7 normalized $\log_2$ expression.

Each endpoint was dichotomized at its cohort median (High = value $\geq$ median), a threshold that is transparent, distribution-free, and yields a near-balanced problem: neutralizing-antibody titer (median = 213; range 20–1280) gave 13 High / 12 Low; $CD8^+$ T-cell activation (median = 4.88; range 0.49–12.1) gave 13 High / 12 Low. The corresponding no-information rate (NIR, the accuracy of always predicting the majority class) was 0.52 for both endpoints, providing the honest baseline against which accuracy is judged. We note (and address in the Discussion) that the neutralizing-antibody median split is partially confounded with trial (Low = 10 Trial-1/2 Trial-2; High = 5 Trial-1/8 Trial-2), whereas $CD8^+$ activation is balanced within each trial and is therefore not trial-confounded.

2.4 Leakage-controlled predictive modeling

Models were evaluated with **Leave-One-Subject-Out Cross-Validation** (LOOCV; 25 folds). Within every fold, using the training subjects *only*, the following pipeline was fit and then applied to the single held-out subject:

1. **Standardization** with `StandardScaler` (mean/variance estimated on the training fold);
2. **Univariate filtering** to the top $k = 20$ genes by absolute point-biserial correlation with the binary target; and
3. an **L1-regularized logistic-regression** classifier (`liblinear`, $C = 1.0$), whose lasso penalty performs additional embedded sparse selection [13, 14].

Because scaling, filtering, and model fitting never observe the held-out subject, the LOOCV estimate is free of the feature-selection leakage that inflates naive cross-validation error [11, 12]. We report AUROC (from the concatenated held-out predicted probabilities), accuracy, sensitivity, specificity, and F_1, each against the NIR baseline, for both endpoints × both feature sets. As an orthogonal generalization test we also trained on Trial 1 ($N = 15$) and tested on Trial 2 ($N = 10$).

2.5 Feature-selection stability

For each model we recorded the set of genes selected in every fold and quantified stability as (i) the mean pairwise Jaccard index across all $\binom{25}{2} = 300$ fold pairs and (ii) the per-gene selection frequency across folds. *Consensus* genes were defined as those selected in $\geq$80% (high-confidence) or 100% of folds. This stability audit is reported alongside performance so that discrimination is never interpreted without knowledge of feature reproducibility.

2.6 Biological interpretation and comparison to YF-17D signatures

For the two best models, the high-confidence ($\geq$80%) consensus genes were subjected to two complementary over-representation analyses (ORA). First, an Enrichr query [18, 19] against Reactome [20], KEGG, WikiPathways, the MSigDB Hallmark collection [21], and GO Biological Process. Second, a background-matched hypergeometric test (`scipy`, Benjamini–Hochberg FDR via `statsmodels`)

against a curated panel of systems-vaccinology signatures—innate/antiviral interferon-stimulated genes (ISGs), ER-stress/unfolded-protein response (UPR), humoral/plasma-cell, T-cell activation, and complement—using the **8,433 measured genes as the assay-aligned background universe** (a critical control that prevents spurious enrichment against a whole-genome background). Finally, every consensus gene was audited for membership in the curated YF-17D signatures drawn from Querec et al. [8] and Gaucher et al. [7].

3 Results

3.1 Cohort and study design

The curated cohort comprises 25 YF-17D–vaccinated healthy adults with paired Day 0 and Day 7 blood transcriptomes and two per-subject adaptive-immune endpoints (Table 1). Sampling density peaks at Days 0, 3, and 7 (25 subjects each), with sparser Day 1 and Day 21 coverage confined to Trial 1 (Fig. 2, left). Neutralizing-antibody titers spanned a ∼64-fold range (20–1280; median 213) and $CD8^+$ T-cell activation spanned ∼25-fold (0.49–12.1; median 4.88), providing adequate dynamic range for a median dichotomization (Fig. 2, right). A principal-component analysis of the gene-level matrix revealed only a mild trial separation along PC1 and no dominant time-point structure in the top components (Fig. 3), consistent with a modest batch effect that the within-subject LFC representation is designed to absorb.

Table 1: Curated YF-17D cohort (NCBI GEO GSE13485). Both adaptive-immune endpoints are available for all 25 subjects, and both Day 0 and Day 7 are present for every subject.

Trial	**Subjects**	**Days sampled**	**Samples**	**Neut.-Ab titer**	**$CD8^+$ activation**
Trial 1 (discovery)	15	0, 1, 3, 7, 21	57	median 213	median 4.88
Trial 2 (validation)	10	0, 3, 7	30	(20–1280)	(0.49–12.1)
Total	**25**	—	**87**	13 High / 12 Low	13 High / 12 Low

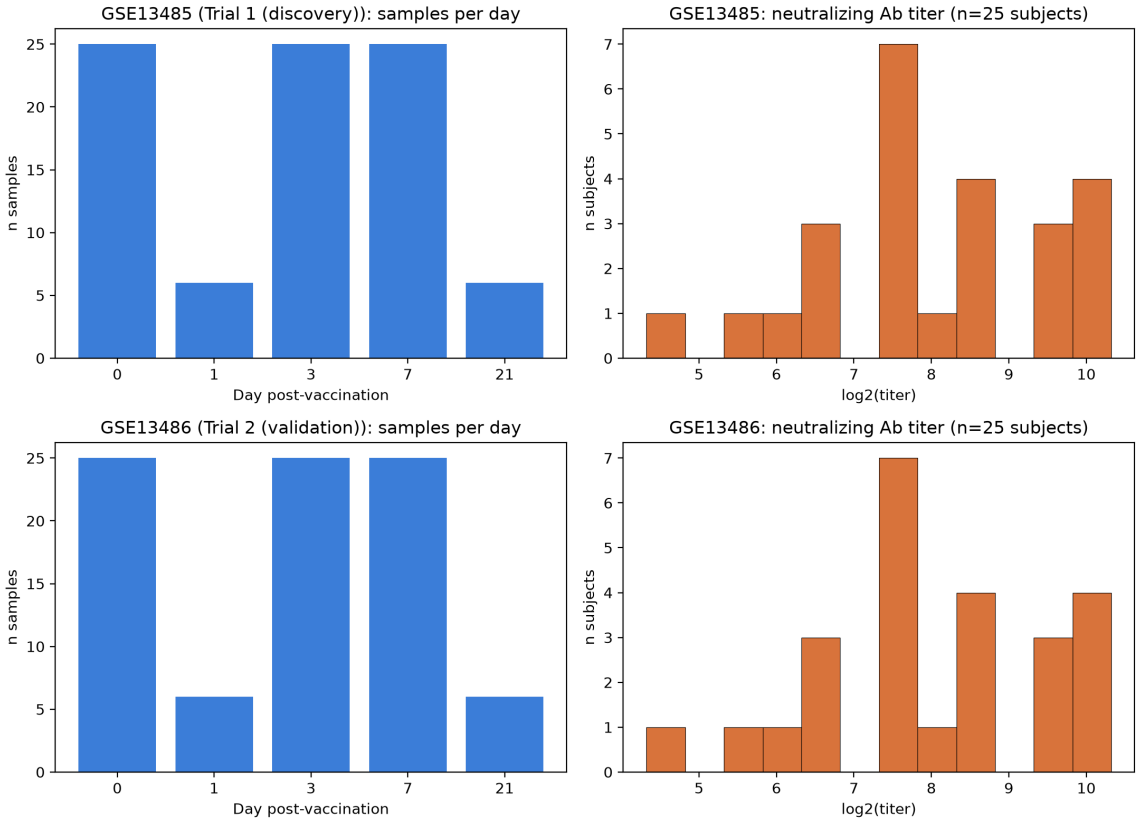

Figure 2: Cohort composition. Left: number of blood samples per post-vaccination day for each trial (GSE13485 = Trial 1; the GSE13486 GPL7567 record is a duplicate of the same arrays). Right: distribution of neutralizing-antibody titers ($\log_2$ scale) across the 25 subjects. The two rows depict the same 25-subject cohort as hosted under the two accession numbers, confirming the SuperSeries duplication documented in Methods.

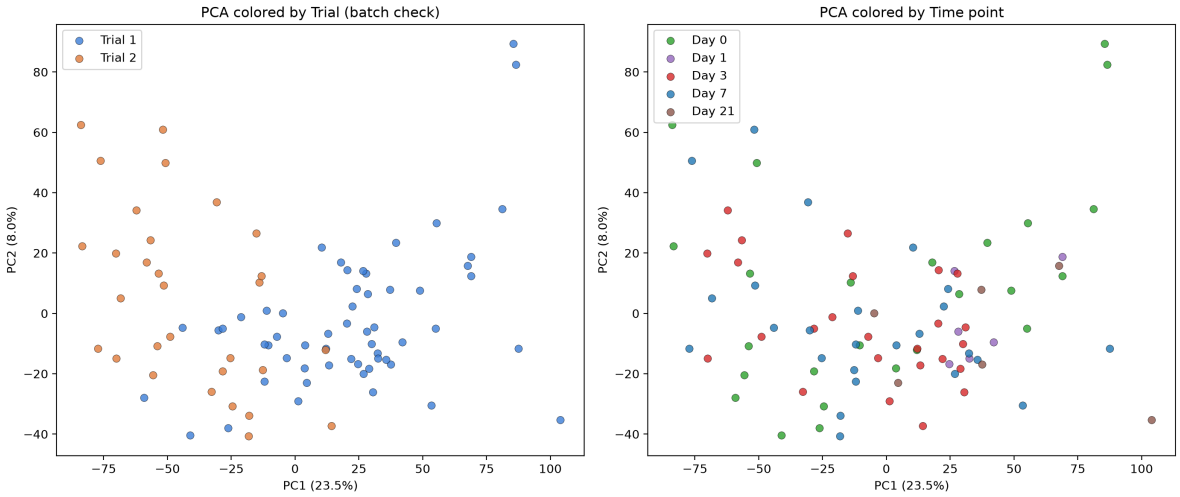

Figure 3: Batch and time-point structure. Principal-component analysis of the 8,433-gene expression matrix (all 87 samples). Left: colored by trial, showing a mild Trial 1/Trial 2 separation along PC1 (23.5% variance). Right: colored by post-vaccination day, showing no dominant time-point axis in the leading components. The within-subject Day 7 − Day 0 $\log_2$ fold-change representation is used to absorb the residual trial offset.

3.2 Early transcriptomics predicts both adaptive-immune endpoints above chance

Under strictly nested LOOCV, both endpoints were predicted above the no-information rate of 0.52, and each was best served by a different feature representation (Table 2; Fig. 4). The **neutralizing-antibody** class was best predicted from the baseline-corrected **LFC** features, reaching an LOOCV AUROC of **0.87** with 0.80 accuracy (20/25 subjects correct), 0.77 sensitivity, and 0.83 specificity. The **CD8$^+$ T-cell activation** class was instead best predicted from **raw Day 7** expression, with an AUROC of **0.76** and 0.72 accuracy (18/25 correct). For each endpoint the alternative feature set performed at or near chance (neutralizing-antibody Day7 AUROC = 0.54; CD8$^+$ LFC AUROC = 0.64), indicating that the informative representation is endpoint-specific: an antibody signal that is carried by the *change* from baseline, and a CD8$^+$ signal that is carried by the *Day-7 state*.

Table 2: Leakage-controlled LOOCV performance (25 folds) for both endpoints × both feature sets. Best model per endpoint in **bold**. All metrics are computed on held-out subjects; accuracy should be read against the no-information rate (NIR = 0.52).

Endpoint	Features	AUROC	Acc.	Sens.	Spec.	F_1	NIR	Cross-trial AUROC
Neut.-Ab titer	**LFC**	**0.87**	**0.80**	0.77	0.83	0.80	0.52	1.00†
	Day7	0.54	0.52	0.54	0.50	0.54	0.52	0.38
CD8$^+$ activation	LFC	0.64	0.52	0.62	0.42	0.57	0.52	0.48
	Day7	**0.76**	**0.72**	0.69	0.75	0.72	0.52	0.56

† The perfect cross-trial AUROC for the antibody model partly reflects the trial/batch axis (the median split is trial-confounded; see §4.2); the within-subject LFC LOOCV AUROC of 0.87 is the trustworthy estimate.

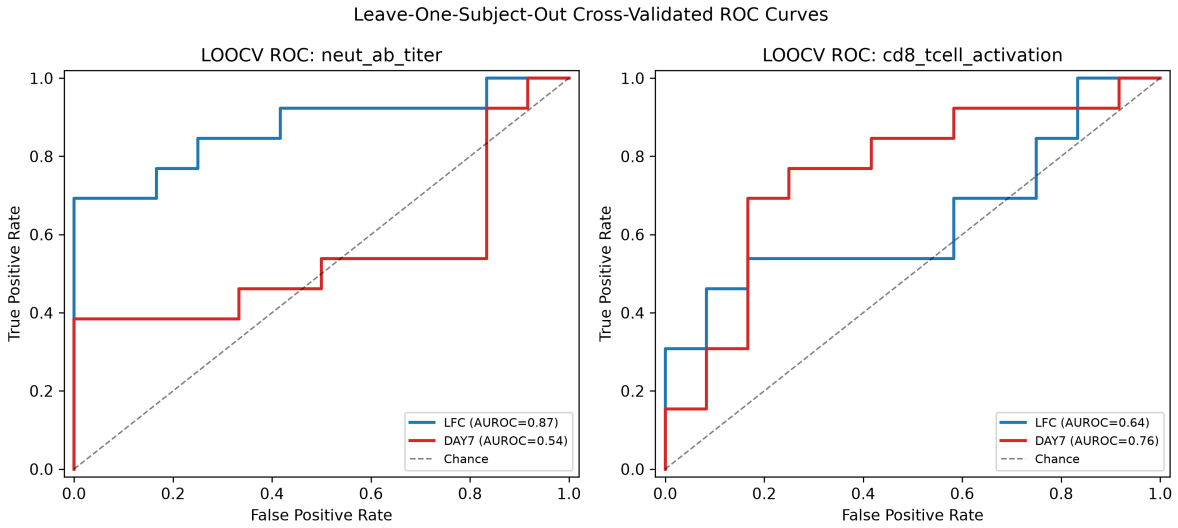

Figure 4: LOOCV ROC curves. Left: neutralizing-antibody titer, where the baseline-corrected LFC features (blue; AUROC = 0.87) clearly outperform raw Day 7 expression (red; AUROC = 0.54). Right: $CD8^+$ T-cell activation, where the ordering reverses and raw Day 7 expression (red; AUROC = 0.76) outperforms LFC (blue; AUROC = 0.64). Diagonal: chance.

The cross-trial (Trial 1→Trial 2) analysis corroborated a genuine antibody signal but also exposed a confound: the antibody LFC model attained a perfect cross-trial AUROC of 1.00, yet because the median antibody split is partially aligned with trial (Trial 1 titers top out where Trial 2 titers begin), part of this separation reflects the batch axis rather than pure vaccine biology. We therefore treat the leakage-free *LOOCV* AUROC of 0.87—which mixes subjects across trials in every fold—as the trustworthy estimate for the antibody endpoint. The $CD8^+$ endpoint, which is balanced within each trial, is not subject to this confound; its more modest cross-trial AUROC (0.56) is consistent with a real but weaker and less trial-portable signal.

3.3 Feature selection is moderately stable and yields compact consensus gene sets

Because discrimination alone cannot establish that a "signature" is reproducible, we quantified the stability of the selected genes across the 25 folds (Fig. 5). Mean pairwise Jaccard overlap between fold-specific top-20 gene sets was moderate for all models (0.48–0.55), with individual fold pairs ranging from ∼0.18 to ∼0.90. Despite this fold-to-fold churn, each model produced a compact core of genes selected in nearly every fold. For the best antibody model (LFC), 11 genes reached ≥80% selection frequency and 4 were selected in 100% of folds (*TGM4*, *GRIN2D*, *PCLAF*, *SLC6A3*). For the best $CD8^+$ model (Day7), 10 genes reached ≥80% and 9 were selected in every fold (*SCGB1D1*, *ZNF132*, *IL5*, *CEP112*, *GALR1*, *KLF17*, *DRC4*, *LINC01465*, *MPHOSPH8*). The full consensus lists are given in Table 3. The moderate Jaccard values are important context: they indicate that although a stable core exists, the majority of the top-20 features in any given fold are interchangeable, a hallmark of feature selection in the large-P, small-N regime [11, 12].

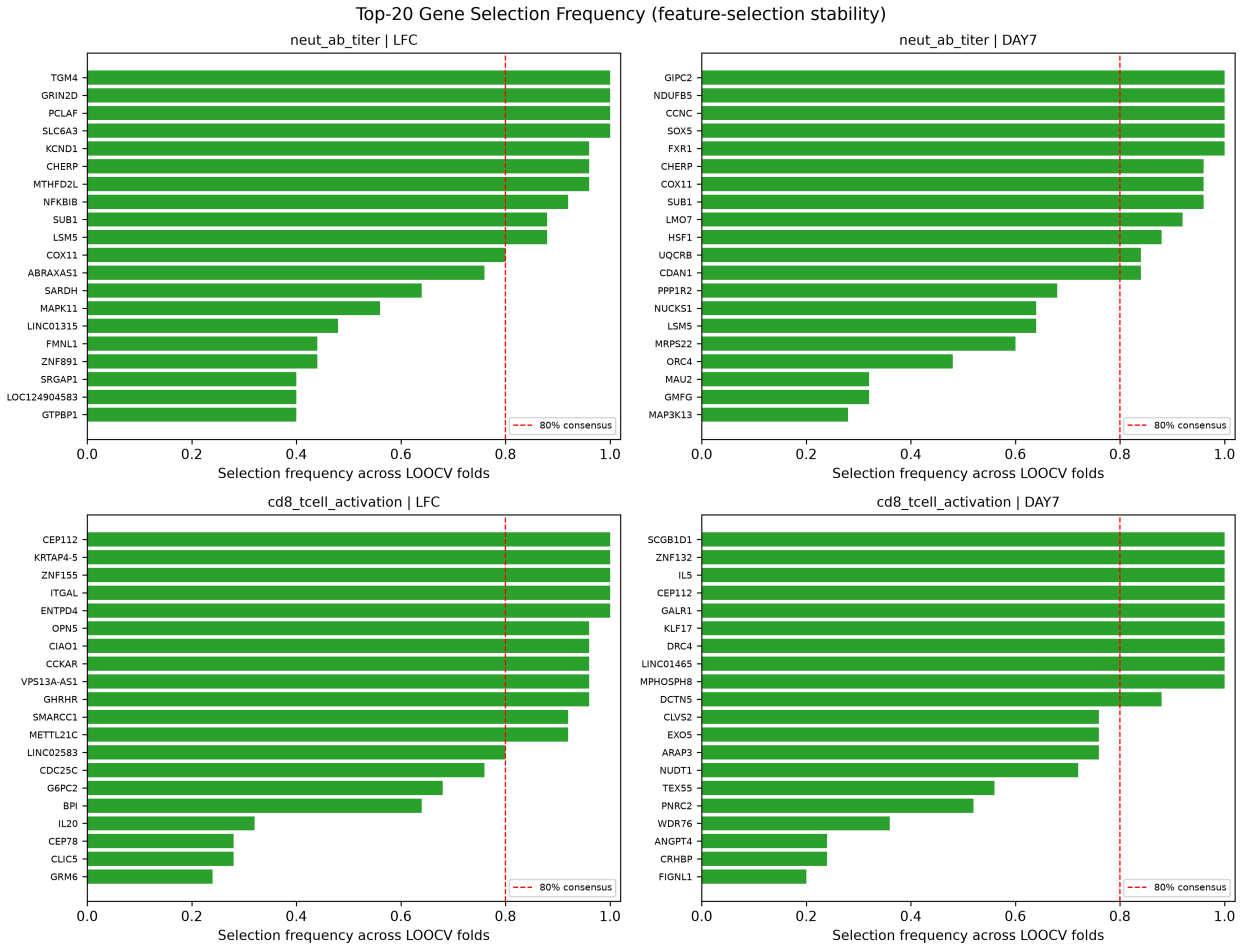

Figure 5: Feature-selection stability. Top-20 gene selection frequency across the 25 LOOCV folds for each endpoint × feature set. The red dashed line marks the 80% consensus threshold. A small core of genes (bars extending to the right of the line) is selected in nearly every fold, but many features fall below the threshold, quantifying the fold-to-fold instability expected in a high-dimensional, low-N problem.

Table 3: High-confidence consensus genes (selected in ≥80% of LOOCV folds) for the two best models, with membership in curated YF-17D signatures. Only *IL5* maps to a canonical module.

Model	**Consensus genes (≥80% of folds)**	**YF-17D signature hits**
Neut.-Ab titer (LFC) AUROC = 0.87	*TGM4*, *GRIN2D*, *PCLAF*, *SLC6A3*, *KCND1*, *CHERP*, *MTHFD2L*, *NFKBIB*, *SUB1*, *LSM5*, *COX11*	none (0/11)
$CD8^+$ activation (Day7) AUROC = 0.76	*SCGB1D1*, *ZNF132*, *IL5*, *CEP112*, *GALR1*, *KLF17*, *DRC4*, *LINC01465*, *MPHOSPH8*, *DCTN5*	*IL5* (T-cell activation)

3.4 Consensus predictive genes dissociate from canonical YF-17D antiviral biology

We next asked whether the consensus predictive genes recapitulate the well-established YF-17D response modules—the type-I interferon/antiviral ISG program, the ER-stress/UPR module, and the humoral plasma-cell program—reported by Querec et al. [8] and Gaucher et al. [7]. The answer is a clear negative (Fig. 6). In the background-matched hypergeometric test against curated signatures, of the 21 high-confidence consensus genes across both best models, **only *IL5*** (a T-helper-2/eosinophil cytokine, in the CD8$^+$ Day7 list) mapped to any canonical module (curated T-cell-activation ORA, adjusted $p = 0.041$). *No* consensus gene overlapped the interferon/ISG (e.g. *STAT1*, *IRF7*, *OAS1/2/3*, *MX1/2*, *ISG15*), ER-stress/UPR (e.g. *DDIT3*, *HERPUD1*, *XBP1*), or plasma-cell/*TNFRSF17* modules that dominate the published YF-17D signal.

Unsupervised Enrichr ORA reinforced this dissociation. For the antibody (LFC) list, the top enriched terms were dominated by ion-transport and neuronal ontologies driven by *SLC6A3*, *GRIN2D*, and *KCND1* (monoatomic cation transmembrane transport, adjusted $p = 0.038$; KEGG "cocaine/amphetamine addiction," adjusted $p = 0.013$), together with metabolic/signaling terms of greater immunological plausibility—MSigDB Hallmark *mTORC1 signaling* (*MTHFD2L*, *NFKBIB*; adjusted $p = 0.010$) and a spliceosome component (*LSM5*, *CHERP*)—none of which are canonical humoral effectors. For the CD8$^+$ (Day7) list, the significant terms were *IL5*-driven interleukin-5 signaling and podosome-assembly ontologies (adjusted $p = 0.024$) alongside *MPHOSPH8*-driven DNA-methylation/heterochromatin terms. The complete ORA table (311 rows) and the per-gene signature-membership audit are provided as supplementary results.

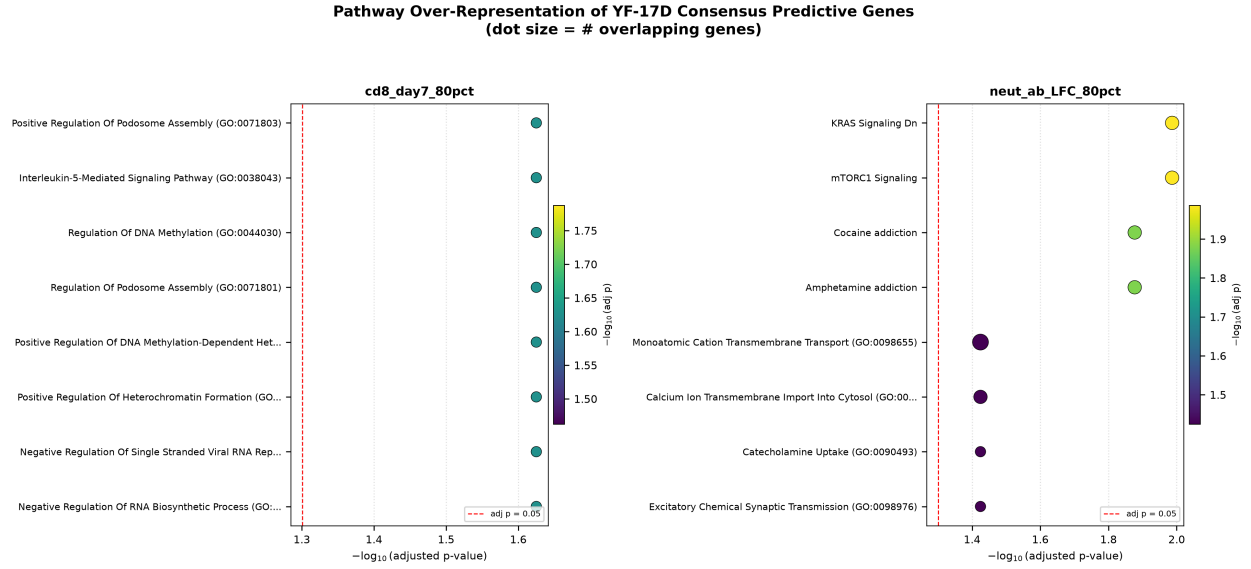

Figure 6: Pathway over-representation of consensus predictive genes. Dot plots of the top enriched terms for the CD8$^+$ Day7 consensus list (left) and the neutralizing-antibody LFC consensus list (right); x-axis = $-\log_{10}$ adjusted p-value, dot size = number of overlapping genes, red dashed line = adjusted-p 0.05 threshold. The enriched terms are dominated by ion-transport/neuronal, metabolic (mTORC1), and epigenetic ontologies rather than the interferon/antiviral, ER-stress, or plasma-cell programs canonical to YF-17D. The only canonical-immunology hit is *IL5*-driven interleukin-5 signaling in the CD8$^+$ list.

3.5 Per-subject predictions

To make the model behaviour fully transparent, Table 4 reports the held-out LOOCV prediction for every subject under both best models. The antibody model errs on 5/25 subjects, concentrated among mid-range titers near the median threshold (e.g. T1_1902, T1_1912, T2_E08), consistent with the intrinsic difficulty of a median dichotomization for subjects close to the split point. The $CD8^+$ model errs on 7/25 subjects, distributed across both trials. Predicted probabilities (not just hard calls) are provided so that decision thresholds other than 0.5 can be examined; the complete machine-readable prediction tables, including the cross-trial predictions, accompany the code.

Table 4: Per-subject held-out LOOCV predictions for the two best models. For each endpoint: $\mathbf{y}$ = true class (1 = High, 0 = Low), $\mathbf{p}$ = predicted probability of High, $\hat{\mathbf{y}}$ = predicted class, and a check (✓)/cross (✗) for correct/incorrect. Neut.-Ab uses the LFC model (AUROC 0.87); $CD8^+$ uses the Day7 model (AUROC 0.76).

		Neut.-Ab titer (LFC)				**$CD8^+$ activation (Day7)**			
Subject	**Trial**	**y**	**p**	**$\hat{y}$**		**y**	**p**	**$\hat{y}$**	
T1_1901	1	1	0.88	1	✓	1	0.58	1	✓
T1_1902	1	0	0.63	1	✗	1	0.96	1	✓
T1_1903	1	0	0.00	0	✓	1	0.82	1	✓
T1_1904	1	0	0.01	0	✓	1	0.96	1	✓
T1_1906	1	0	0.44	0	✓	1	0.82	1	✓
T1_1907	1	0	0.13	0	✓	1	0.23	0	✗
T1_1908	1	0	0.37	0	✓	0	0.45	0	✓
T1_1909	1	1	0.19	0	✗	1	0.46	0	✗
T1_1910	1	1	0.87	1	✓	0	0.00	0	✓
T1_1911	1	1	0.74	1	✓	0	0.10	0	✓
T1_1912	1	0	0.72	1	✗	0	0.12	0	✓
T1_1913	1	0	0.02	0	✓	0	0.29	0	✓
T1_1914	1	0	0.02	0	✓	0	0.15	0	✓
T1_1915	1	1	1.00	1	✓	1	0.70	1	✓
T1_1916	1	0	0.24	0	✓	0	0.90	1	✗
T2_B02	2	0	0.17	0	✓	1	0.36	0	✗
T2_B03	2	1	0.92	1	✓	0	0.35	0	✓
T2_B08	2	1	0.95	1	✓	0	0.13	0	✓
T2_B09	2	1	0.91	1	✓	1	0.63	1	✓
T2_D03	2	1	0.60	1	✓	0	0.39	0	✓
T2_E04	2	1	0.01	0	✗	1	0.03	0	✗
T2_E08	2	1	0.41	0	✗	1	0.65	1	✓
T2_E09	2	0	0.15	0	✓	0	0.78	1	✗
T2_G01	2	1	0.99	1	✓	0	0.52	1	✗
T2_G03	2	1	0.78	1	✓	1	0.52	1	✓
Correct / 25		**20 (0.80)**				**18 (0.72)**			

4 Discussion

Re-analysing the canonical YF-17D cohort with a leakage-controlled machine-learning workflow yields two findings that must be held together. First, early post-vaccination blood transcriptomics carries a *genuine, cross-validated* predictive signal for adaptive-immune outcomes: under strictly nested LOOCV, neutralizing-antibody class is predicted with AUROC 0.87 from the Day 7 − Day 0 fold-change, and $CD8^+$ T-cell activation with AUROC 0.76 from Day 7 expression, both comfort-

ably above the 0.52 no-information rate. This affirms, with modern methodological guardrails, the core premise of systems vaccinology as originally advanced for YF-17D [1, 6, 8]. Second, and more soberingly, the specific genes that drive this prediction *dissociate* almost entirely from the canonical YF-17D biology: of 21 high-confidence consensus genes, only *IL5* maps to an established immune module, and none touch the interferon/ISG, ER-stress/UPR, or plasma-cell programs that Gaucher et al. [7] and Querec et al. [8] showed to dominate the YF-17D response.

4.1 Why predictive genes need not be mechanistic drivers

We interpret this dissociation not as a refutation of the known YF-17D biology but as a textbook consequence of feature selection in the high-dimensional (P = 8,433), low-sample-size (N = 25) regime. When thousands of correlated features compete to explain a binary label observed in only 25 subjects, a sparse L1-penalized classifier will select whichever small set of probes happens to be most discriminative in the training fold—and, because many genes carry partially redundant information, that set need not coincide with the biologically dominant response modules [13, 14]. Our own stability analysis makes this concrete: mean fold-to-fold Jaccard overlap was only ∼0.48–0.55, meaning roughly half of the top-20 features are interchangeable between folds even as the model predicts well. This is precisely the behaviour anticipated by the foundational cautions of Ambroise and McLachlan [11] and Varma and Simon [12], who showed that feature sets extracted from microarray data are unstable and that only rigorously nested cross-validation yields honest error estimates. The consensus genes we report should therefore be read as *predictive markers under this design*, not as validated mechanistic drivers, and biological narratives built on individual genes such as *TGM4* or *SLC6A3* would be unwarranted.

It is nonetheless notable that a minority of consensus genes are immunologically coherent. *IL5*, the sole canonical hit, is a T-helper-2/eosinophil cytokine consistent with the balanced Th1/Th2 program that YF-17D induces via multi-TLR dendritic-cell activation [5]. Among the antibody-model genes, *NFKBIB* (an NF-κB inhibitor) and the Hallmark-mTORC1 members *MTHFD2L* and *NFKBIB* implicate NF-κB and mTORC1 signaling axes that are plausibly linked to lymphocyte activation and antibody-secreting-cell differentiation. These are hypotheses, not conclusions; the dominant enriched ontologies (ion transport, neuronal signaling) are far more likely to reflect selection variance than vaccine physiology.

4.2 The trial confound and the value of baseline correction

A specific caveat attends the antibody endpoint. Because Trial 1 and Trial 2 differ systematically in their titer distributions, the median split is partially confounded with trial, so the perfect cross-trial AUROC (1.00) cannot be attributed to vaccine biology alone. We deliberately foreground the LOOCV estimate (0.87)—which randomizes subjects across trials in every fold and is therefore not driven by a single batch axis—as the defensible performance figure, and we note that the $CD8^+$ endpoint, balanced within trials, is free of this confound. The finding that antibody prediction favours the *within-subject* LFC representation while $CD8^+$ prediction favours the *raw Day 7* state is itself informative: baseline correction removes subject- and trial-specific offsets that would otherwise dominate, and its benefit for the trial-confounded antibody endpoint is consistent with the mild batch structure seen in the PCA (Fig. 3).

4.3 Relation to the broader systems-vaccinology literature

Our results sit comfortably within, and add nuance to, a decade of systems-vaccinology work. Predictive early signatures have been reported for influenza [9] and, in cross-vaccine synthesis, for five human vaccines [10], the latter also establishing the Blood Transcription Modules as a robust, module-level vocabulary that is far less brittle than individual-gene signatures. The contrast between those module-level successes and our gene-level dissociation suggests a concrete methodological lesson: the reproducible, mechanistically interpretable signal in small vaccine cohorts likely lives at the level of *coordinated modules* (interferon, plasma-cell, cell-cycle) rather than individual transcripts, and a classifier free to pick single genes will tend to bypass it. In other words, the canonical YF-17D antiviral signal may well be present in these data at the signature level even though our sparse classifier does not require it to predict.

4.4 Limitations

Several limitations bound our conclusions. (i) The cohort is small ($N = 25$) and single-study; LOOCV, while leakage-free, still yields performance estimates with wide confidence intervals at this sample size, and the SuperSeries duplication means no truly independent external cohort was available here. (ii) The custom cDNA/EST-heavy platform resolved gene symbols for only ∼42% of probes, so 8,433 genes—though ample—do not cover the full transcriptome, and some canonical ISGs may be under-represented on the array. (iii) Median dichotomization discards magnitude information and places borderline subjects at maximal risk of misclassification, as seen in the per-subject errors; continuous-outcome regression is a natural complement. (iv) Expression values are the authors' deposited normalized intensities rather than re-processed raw signals. (v) The biological interpretation rests on over-representation analysis of a thresholded consensus list, which is less powerful than a rank-based enrichment on the full feature set.

4.5 Future directions

Three directions follow directly. First, *independent-cohort validation*: the consensus genes should be tested in a genuinely separate YF-17D cohort before any are advanced as biomarkers. Second, *module-level modeling*: repeating the analysis with Blood Transcription Modules [10] or Hallmark gene sets [21] as features, and applying rank-based Gene Set Enrichment Analysis to the full ranked feature list rather than a thresholded consensus, would test our hypothesis that the canonical interferon/plasma-cell signal is recoverable at the module level even when the sparse gene-level classifier bypasses it. Third, *stability-aware model selection*: incorporating feature-selection stability directly into model choice (e.g. stability selection or ensemble consensus) would help ensure that reported signatures are reproducible, not merely discriminative. Together these steps would sharpen the distinction—central to this work—between predicting an immune outcome and explaining it.

4.6 Conclusion

Early blood transcriptomics predicts YF-17D adaptive-immune outcomes above chance under a rigorous, leakage-free evaluation, confirming the systems-vaccinology premise. But the sparse genes that carry this prediction largely dissociate from the vaccine's canonical antiviral, ER-stress, and plasma-cell biology, illustrating that in the large-P, small-N setting a predictive marker is not the same as a mechanistic driver. Honest reporting of feature-selection stability, transparent per-subject

predictions, and independent validation are therefore prerequisites before any transcriptomic "signature" is interpreted biologically or deployed as a biomarker.

Code and Data Availability

All primary data are publicly available from NCBI GEO under accession GSE13485 (with duplicated GPL7567 arrays under the SuperSeries GSE13486) [16, 17]. The complete analysis pipeline—data download and curation, preprocessing and probe-to-gene collapse, leakage-controlled LOOCV modeling, feature-stability analysis, and pathway overrepresentation analysis—is provided as documented, seeded Python scripts (`download_data.py`, `parse_data.py`, `fetch_probe_mapping.py`, `preprocess_data.py`, `train_models.py`, `biological_interpretation.py`). Per-subject predictions (`predictions_neut_ab.csv`, `predictions_cd8.csv`), full performance metrics (`model_performance_metrics.json`), consensus-gene selection frequencies (`consensus_genes.csv`), and the complete enrichment table (`enrichment_results.csv`) accompany the manuscript.

Acknowledgements

We gratefully acknowledge the investigators of the original YF-17D systems-vaccinology study for depositing their data in the public domain, enabling this re-analysis.

References

[1] Bali Pulendran, Shuzhao Li, and Helder I. Nakaya. Systems vaccinology. *Immunity*, 33(4): 516–529, 2010. doi: 10.1016/j.immuni.2010.10.006.

[2] Thomas Hagan, Helder I. Nakaya, Shankar Subramaniam, and Bali Pulendran. Systems vaccinology: Enabling rational vaccine design with systems biological approaches. *Vaccine*, 33 (40):5294–5301, 2015. doi: 10.1016/j.vaccine.2015.03.072.

[3] Thomas P. Monath. Yellow fever: an update. *The Lancet Infectious Diseases*, 1(1):11–20, 2001. doi: 10.1016/S1473-3099(01)00016-0.

[4] Rama S. Akondy, Nathan D. Monson, Joseph D. Miller, Srilatha Edupuganti, Dirk Teuwen, Haiyan Wu, Farah Quyyumi, Seema Garg, John D. Altman, Carlos Del Rio, Harry L. Keyserling, Alexander Ploss, Charles M. Rice, Walter A. Orenstein, Mark J. Mulligan, and Rafi Ahmed. The yellow fever virus vaccine induces a broad and polyfunctional human memory CD8+ T cell response. *The Journal of Immunology*, 183(12):7919–7930, 2009. doi: 10.4049/jimmunol.0803903.

[5] Troy Querec, Soumaya Bennouna, Sertac Alkan, Yasmina Laouar, Keith Gorden, Richard Flavell, Shizuo Akira, Rafi Ahmed, and Bali Pulendran. Yellow fever vaccine YF-17D activates multiple dendritic cell subsets via TLR2, 7, 8, and 9 to stimulate polyvalent immunity. *Journal of Experimental Medicine*, 203(2):413–424, 2006. doi: 10.1084/jem.20051720.

[6] Bali Pulendran. Learning immunology from the yellow fever vaccine: innate immunity to systems vaccinology. *Nature Reviews Immunology*, 9(10):741–747, 2009. doi: 10.1038/nri2629.

[7] Denis Gaucher, Rafick Therrien, Nadia Kettaf, Bastian R. Angermann, Geneviève Boucher, Abdelali Filali-Mouhim, Janet M. Moser, Rahul S. Mehta, Donald R. Drake, Elias Castro, Rama Akondy, Aline Rinfret, Bader Yassine-Diab, Elias A. Said, Yoav Chouikh, Mark J. Cameron, Robert Clum, David Kelvin, Ron Somogyi, Larry D. Greller, Robert S. Balderas, Peter Wilkinson, Giuseppe Pantaleo, James Tartaglia, Elias K. Haddad, and Rafick-Pierre Sekaly. Yellow fever vaccine induces integrated multilineage and polyfunctional immune responses. *Journal of Experimental Medicine*, 205(13):3119–3131, 2008. doi: 10.1084/jem.20082292.

[8] Troy D. Querec, Rama S. Akondy, Eva K. Lee, Weiping Cao, Helder I. Nakaya, Dirk Teuwen, Ali Pirani, Kim Gernert, Jiusheng Deng, Bruz Marzolf, Kathleen Kennedy, Haiyan Wu, Soumaya Bennouna, Herold Oluoch, Joseph Miller, Ricardo Z. Vencio, Mark Mulligan, Alan Aderem, Rafi Ahmed, and Bali Pulendran. Systems biology approach predicts immunogenicity of the yellow fever vaccine in humans. *Nature Immunology*, 10(1):116–125, 2009. doi: 10.1038/ni.1688.

[9] Helder I. Nakaya, Jens Wrammert, Eva K. Lee, Luigi Racioppi, Stephanie Marie-Kunze, W. Nicholas Haining, Anthony R. Means, Sudhir P. Kasturi, Nooruddin Khan, Gui-Mei Li, Megan McCausland, Vishal Kanchan, Kenneth E. Kokko, Shuzhao Li, Rivka Elbein, Aneesh K. Mehta, Alan Aderem, Kanta Subbarao, Rafi Ahmed, and Bali Pulendran. Systems biology of vaccination for seasonal influenza in humans. *Nature Immunology*, 12(8):786–795, 2011. doi: 10.1038/ni.2067.

[10] Shuzhao Li, Nadine Rouphael, Sai Duraisingham, Sandra Romero-Steiner, Scott Presnell, Carl Davis, Daniel S. Schmidt, Scott E. Johnson, Andrea Milton, Gowrisankar Rajam, Sudhir Kasturi, George M. Carlone, Charlie Quinn, Damien Chaussabel, A. Karolina Palucka, Mark J. Mulligan, Rafi Ahmed, David S. Stephens, Helder I. Nakaya, and Bali Pulendran. Molecular signatures of antibody responses derived from a systems biology study of five human vaccines. *Nature Immunology*, 15(2):195–204, 2014. doi: 10.1038/ni.2789.

[11] Christophe Ambroise and Geoffrey J. McLachlan. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Sciences*, 99 (10):6562–6566, 2002. doi: 10.1073/pnas.102102699.

[12] Sudhir Varma and Richard Simon. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7:91, 2006. doi: 10.1186/1471-2105-7-91.

[13] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996. doi: 10.1111/j.2517-6161.1996.tb02080.x.

[14] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005. doi: 10.1111/j.1467-9868.2005.00503.x.

[15] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[16] Ron Edgar, Michael Domrachev, and Alex E. Lash. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210, 2002. doi: 10.1093/nar/30.1.207.

[17] Tanya Barrett, Stephen E. Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F. Kim, Maxim Tomashevsky, Kimberly A. Marshall, Katherine H. Phillippy, Patti M. Sherman, Michelle Holko, Andrey Yefanov, Hyeseung Lee, Naigong Zhang, Cynthia L. Robertson, Nadezhda Serova, Sean Davis, and Alexandra Soboleva. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research*, 41(D1):D991–D995, 2013. doi: 10.1093/nar/gks1193.

[18] Edward Y. Chen, Christopher M. Tan, Yan Kou, Qiaonan Duan, Zichen Wang, Gabriela V. Meirelles, Neil R. Clark, and Avi Ma'ayan. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*, 14:128, 2013. doi: 10.1186/1471-2105-14-128.

[19] Maxim V. Kuleshov, Matthew R. Jones, Andrew D. Rouillard, Nicolas F. Fernandez, Qiaonan Duan, Zichen Wang, Simon Koplev, Sherry L. Jenkins, Kathleen M. Jagodnik, Alexander Lachmann, Michael G. McDermott, Caroline D. Monteiro, Gregory W. Gundersen, and Avi Ma'ayan. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*, 44(W1):W90–W97, 2016. doi: 10.1093/nar/gkw377.

[20] Bijay Jassal, Lisa Matthews, Guilherme Viteri, Chuqiao Gong, Pascual Lorente, Antonio Fabregat, Konstantinos Sidiropoulos, Justin Cook, Marc Gillespie, Robin Haw, Fred Loney, Bruce May, Marija Milacic, Karen Rothfels, Cristoffer Sevilla, Veronica Shamovsky, Solomon Shorser, Thawfeek Varusai, Joel Weiser, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D'Eustachio. The reactome pathway knowledgebase. *Nucleic Acids Research*, 48(D1):D498–D503, 2020. doi: 10.1093/nar/gkz1031.

[21] Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P. Mesirov, and Pablo Tamayo. The molecular signatures database hallmark gene set collection. *Cell Systems*, 1(6):417–425, 2015. doi: 10.1016/j.cels.2015.12.004.