The Impact of CMV on T Cell Repertoire: Insights from Transcriptomic Data Analysis

Gokul Seshadri
University of Minnesota Twin Cities
sesha059@umn.edu
Advisors: Bharat Thyagarajan and Adam Rothman

April 02, 2025

Abstract

Cytomegalovirus (CMV) is a very prevalent virus and was implicated widely to be a risk factor of all cause mortality and can make T-cell receptors (TCR) to be CMV specific, which might hinder the diversity of T-cell repertoire. Hence in this study, we have investigated the impact of CMV on T-cell repertoire and their contributions to mortality risk using gene expression pathways. Previous research identified the percentage of CD4+ naïve T cells (CD4N) as a T-cell aging marker associated with decreased biological aging and reduced mortality risk, hence we have included CD4N as a component of T-cell repertoire. In the HRS older adults, we found that CMV drives the expansion of TCR clones (ATE: 8.5 [4.5 to 12.5]) but had a detrimental effect on the percentage of CD4N (ATE: -6.6 [-8.1 to -5.2]), despite both TCR clones and CD4N being protective of all-cause mortality. To further investigate this paradox, we examined genes that mediate the relationship between CMV infection and T cell repertoire. Our analysis using gene co-expression modules and latent modeling revealed that CMV impacts T cell alpha-beta clones and CD4N through cell trafficking, adhesion, and activation pathways, and among them the cell trafficking pathway was shown to be a strong contributor toward the decreased mortality association of the T-cell repertoire, and this finding was replicated in an independent cohort of individuals from the Long-Life Family Study. Thus, our findings illustrate that a certain pathway could serve as a promising target for immunological interventions to mitigate the detrimental effects of CMV on the repertoire of T cells.

1 Introduction

The human immune system plays a multifaceted role in safeguarding our body against many foreign substances. One such role is adaptive immunity, also known as acquired immune response. The adaptive immune system is highly pathogen-specific: T cells and B cells are the major players in the adaptive immune system. They recognize specific “non-self” antigens, generate tailored immune responses to eliminate specific pathogens and develop immunological memory so that they can later induce secondary immune responses upon renewed contact with the same pathogen. T cells are classified based on the specific markers they express on their surface and their functional roles in the immune system. The two main types are helper T cells (CD4+ T cells) and cytotoxic T cells (CD8+ T cells). Helper T cells (CD4+) coordinate immune responses by activating other immune cells, such as B cells and cytotoxic T cells, and by secreting cytokines. Cytotoxic T cells (CD8+) directly kill infected or cancerous cells by recognizing antigens presented by MHC class I molecule[1].

T cell receptors (TCRs)[2] are protein complexes on the surface of T cells that identify and bind to foreign substances, helping the body fight infections. They are generated through a process called V(D)J recombination, where gene segments encoding the TCR α, β, γ, and δ chains are randomly arranged in developing T cells within the thymus. The majority (95%) of TCRs consist of α (TRA) and β (TRB) chains which recognize antigens bound to major histocompatibility complex (MHC molecules), there is also a small set (5%) of γ and δ chains which recognize antigens without MHC presentation. The predominance of $\alpha\beta$ TCR clones in humans is due to their greater combinatorial diversity and their combinations enable diverse and precise antigen recognition, balancing specificity and diversity to target pathogens effectively. A T-cell clone is a group of T cells that all share an identical T-cell receptor (TCR) sequence, meaning they originate from a single parent T cell that underwent a unique V(D)J gene rearrangement during its development in the thymus[3]. This unique TCR allows each clone to recognize a specific antigen. When a naive T cell encounters its matching antigen, it becomes activated and undergoes clonal expansion, rapidly producing many identical daughter cells that can mount an effective immune response against the pathogen. After the immune response, most of these effector cells die, but some persist as memory T cells, ready to respond quickly if the same antigen is encountered again. The diversity and abundance of T-cell clones are crucial for the immune system’s ability to recognize a wide array of pathogens and are fundamental to adaptive immunity[4]. T cells have been studied extensively from different facets including quantifying the changes in the collection of unique T cell receptors under different pathological conditions[5, 6]. Also, previous studies have shown that T cell subsets like CD4+ naïve T cells[7, 8] were significantly associated with reduced mortality risk.

Cytomegalovirus (CMV) is a chronic viral infection and is a risk factor for mortality for those with a weakened immune system[9]. In older adults, CMV tends to increase the proportion of senescent T cells[10], which is characterized by decreased production of interleukin-2 and less dependence on costimulation. CMV also causes clonal expansion of CD8+ T cells (cytotoxic)[11]; a significant portion of the T cell repertoire is sometimes CMV-specific, which may reduce the diversity of the T cell repertoire or alter the cellular functions, potentially impairing the ability to respond to new infections. Thus, CMV infection can skew the T cell response, prioritizing CMV-specific clones over other potential immune challenges[12], and this is particularly concerning in older individuals, where immune resources are already limited. Hence, understanding the effects of CMV infection on the T cell clones is vital for designing strategies to mitigate CMV's long-term impact on health.

In an earlier study[5], authors have profiled the repertoire of 666 individuals with known CMV serostatus and have measured the association of each clonotype with CMV seropositivity using Fisher's exact test, and this study enabled inference of CMV serostatus from the profiled TCR-β clonotypes. In recent years, machine learning methods were used to identify clonotypes that are predictors of immunotherapy response[13] and predict the interaction between a TCR sequence and an unknown antigen specificity[14]. Despite extensive research on T cell receptors (TCRs), to the best of our knowledge, no prior study has examined the broader impact of chronic viral infections like cytomegalovirus (CMV) on the overall diversity and adaptability of the T-cell immune response.

Therefore, in this study, we investigated the impact of a common viral infection like CMV on the T-cell repertoire and their contributions to mortality risk through gene expression pathways. While the repertoire typically refers to T-cell receptors alone, we also incorporated the percentage of CD4 naïve T cells—previously shown to be associated with age related diseases, biological aging and mortality[8]—to better capture the adaptability of T cells in responding to antigens. Utilizing bulk RNA-sequencing data from the U.S. Health and Retirement Study, we applied the MiXCR pipeline to extract TCR clonotypes from both α and β chains. Furthermore, we leveraged gene expression profiles to identify major biological pathways that mediate the relationship between CMV infection and the T-cell repertoire, and evaluated which of these pathways may contribute to reduced mortality risk associated with a more diverse and adaptable T-cell immunity.

2 Methods

2.1 Cohorts

We utilized data from two independent cohorts: the Health and Retirement Study (HRS) served as our discovery cohort, while the Long-Life Family Study (LLFS) functioned as the validation cohort.

The Health and Retirement Study (HRS) is a nationally representative, ongoing panel study of older U.S. adults that began in 1992[15]. As part of the 2016 data collection, venous blood samples were obtained from a subsample of approximately 4,000 participants, with 2.5 mL collected in PAXgene tubes. Total RNA was extracted using the QIACube semi-automated system in combination with the PAXgene Blood miRNA Kit, utilizing 200–500 ng of RNA per assay. RNA was extracted from only half of each PAXgene tube to preserve remaining material in various formats for future use. Ribosomal and globin RNA depletion was performed using the TruSeq Stranded Total Library Prep Gold Kit (Ribo-Zero Gold), and RNA sequencing was carried out on the Illumina NovaSeq platform, generating 50 bp paired-end reads at a minimum depth of 20 million reads per sample. RNA-Seq was successfully completed for 3,685 participants. The RNA-Seq pipeline largely followed the TOPMed/GTEx protocols, with minor modifications. Further details on the RNA-Seq processing pipeline can be found on the HRS website.

T cell subsets like the CD4+ naive T-cells in HRS have been measured using methods based on a standardized protocol published by the Human Immunology Project[16]. Cytomegalovirus (CMV) serostatus was assessed using quantitative IgG antibody measurements in serum, conducted on the Roche e411 immunoassay analyzer. For this analysis, results were dichotomized into positive (reactive) and negative (including non-reactive and borderline cases). Demographic variables like chronological age (in years), sex (female/-male), and race/ethnicity (Hispanic White, Non-Hispanic White, Non-Hispanic Black, and Non-Hispanic Other) were obtained from the HRS demographics dataset. Body Mass Index (BMI) was calculated from height and weight measurements collected during the 2014 and 2016 physical examination waves. Smoking status was self-reported and categorized as never, former, or current smokers. Mortality status was determined during the 2020 interview cycle (4-year mortality) based on reports from informants regarding the death of HRS participants.

The comorbidity index was constructed by summing self-reported chronic conditions from the 2016 HRS survey, which comprises of hypertension, type II diabetes, cancer, lung disease, cardiac disorders, stroke, arthritis, and psychiatric disorders. High-sensitivity C-reactive protein (CRP) levels were measured in serum using a latex-enhanced immunoturbidimetric assay, read on the Roche COBAS 6000 Chemistry analyzer. Inflammatory biomarkers (IL-1RA, IL-6, IL-10, and sTNFR-1) were quantified in serum using Simple Plex assays on the ELLA System. A latent construct representing systemic inflammation was derived through confirmatory factor analysis in MPlus v8, based on log-transformed values of CRP, IL-1RA, IL-10, sTNFR-1, and IL-6.

The Long Life Family Study (LLFS) is a longitudinal cohort comprising nearly 5,000 individuals from 539 families selected for their exceptional longevity[17]. Data have been collected across three waves, approximately 6–8 years apart. Blood samples were collected during the first two waves, and for this study, we utilized data from the first wave to align temporally with the Health and Retirement Study (HRS). More details about the study design are available on the LLFS website. In the LLFS, the number of participant deaths within a 4-year follow-up was relatively low compared to the HRS cohort. Therefore, we extended the follow-up period and used 8-year mortality for the validation analysis. Mortality status was determined based on whether the recorded date of death occurred on or before December 31, 2014.

For Visit 1, RNA sequencing was performed using RNA extracted from PAXgene™ Blood RNA tubes and processed with the Qiagen PreAnalytiX PAXgene Blood miRNA Kit. Library preparation, quality control, and sequencing were conducted by the Division of Computational & Data Sciences at Washington University. The nf-core/rnaseq v3.14.0 pipeline was used for read alignment, duplicate marking, and transcript quantification. Genes with low expression (fewer than 4 counts per million in at least 98.5% of samples) and those with significant intergenic overlap were excluded. The final dataset included 1,810 samples and 16,418 genes. For downstream analyses, raw counts were normalized and transformed to the log2 counts per million (log2CPM) scale.

2.2 MiXCR

MiXCR is a fast and accurate next-generation sequencing (NGS) software tool designed for comprehensive analysis of T-cell receptor (TCR) and B-cell receptor (BCR) repertoires[18]. It processes raw sequencing reads to identify and quantify unique clonotypes—combinations of V(D)J gene segments and CDR3 regions that define individual immune cell clones. MiXCR performs alignment of reads to germline V, D, J, and C gene segments, assembles clonotypes by grouping reads with identical CDR3 sequences, and applies error correction methods to handle sequencing artifacts and PCR errors. It supports data from various NGS platforms like Illumina and Ion Torrent and generates detailed output files containing information on clonotype abundance, gene usage, and CDR3 sequences. MiXCR is widely used in immunology and cancer research for applications such as tracking clonal expansion, studying immune diversity, and evaluating responses to infections or immunotherapies.

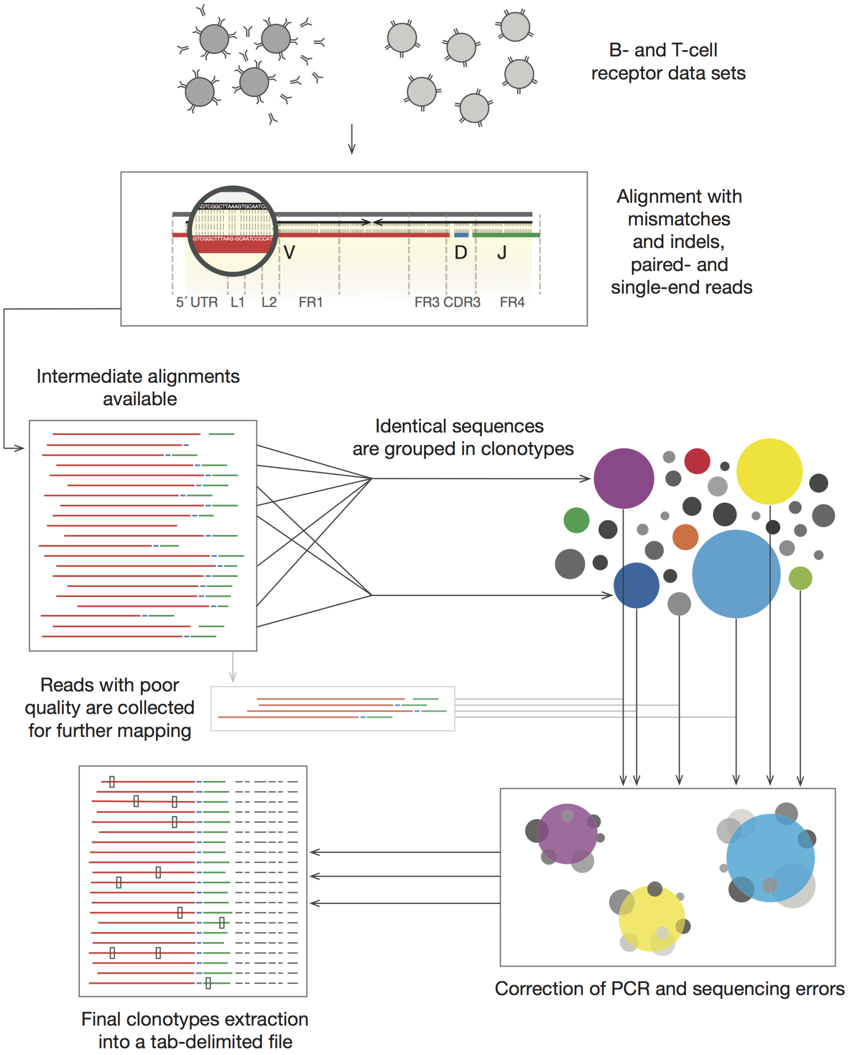

B- and T-cell receptor data sets
Alignment with mismatches and indels, paired- and single-end reads
V
D
J
5´UTR
L1
L2
FR1
FR3
CDR3
FR4
Intermediate alignments available
Identical sequences are grouped in clonotypes
Reads with poor quality are collected for further mapping
Correction of PCR and sequencing errors
Final clonotypes extraction into a tab-delimited file

Figure 1: Flow Diagram of rna-seq pipeline

MiXCR has a set of libraries called presets, which contains a list of pre-configured steps needed to run an analysis for a particular data type. A preset can be invoked with *mixcr analyze* command, and *rna-seq* preset can be used to extract TCR repertoire.

```
Command Snippet:
mixcr analyze rna-seq --species hsa {R1} {R2} {output}
```

The below are the steps involved in the mixcr rna-seq preset,

- **align**: map against V-, D-, J- and C- gene segments
- **assemblePartial**: assembles alignments that only partially cover CDR3 region.
- **extend**: Imputes missing nucleotides
- **assemble**: assembles into clonotypes and applies errors correction.
- **export**: Exports into tsv files

3 Statistical Analysis

Next generation sequencing software MiXCR was used to extract T cell clones from TCR alpha and beta chains, and then clonal properties—total number of clones (total) and the number of unique clones—were calculated. We then ran multivariate logistic regression models to evaluate the association of these clonal properties with mortality after adjusting for age, sex, race, education, smoking status, CMV status, BMI, inflammation, and comorbidity index. The models were also controlled for total read count from the gene expression data as a measure of sequencing depth.

The clonal properties of alpha and beta chains were then combined into a single feature (TRAB) using Principal Component Analysis (PCA). The effect of CMV on T-cell repertoire, TRAB and CD4 naïve T cells, was evaluated using propensity-score-based methods, with propensity scores computed via logistic regression models using CMV as the outcome and age, sex, race, education, smoking status, total T cells, inflammation, and comorbidity index as predictors.

To investigate the pathways through which CMV impacts the T-cell repertoire, we used a series of RNA-seq analysis methods. DESeq2 was used to identify genes differentially expressed between CMV positive and negative groups, WGCNA was used to find co-expression modules among those genes, and Group Lasso was applied to identify gene modules that are strong predictors of TRAB and CD4 naïve T cells. Gene ontology (GO) enrichment analysis was then performed on the resultant gene modules.

To identify the most influential pathways for mortality risk, a deep neural network with sub-networks (encoders) was used to obtain latent representations of each gene module, and this model was trained to predict mortality. While the final layer output provides the cumulative impact of the T-cell repertoire on mortality, the outputs from each subnetwork correspond to the impact of individual gene modules, thereby reflecting the biological pathways involved.

The following subsections outline the key steps of our analysis, accompanied by brief descriptions of the statistical methods used.

3.1 Combining TRA and TRB clonal properties using PCA

Principal Component Analysis (PCA) is a statistical technique used to reduce the dimensionality of complex datasets by transforming variables into a smaller set of uncorrelated components called principal components. These components are linear combinations of the original variables, ordered by their ability to explain variance in the data, with the first component capturing the largest possible variance. By projecting data onto orthogonal axes aligned with directions of maximum variability, PCA simplifies datasets while preserving their essential structure. It is particularly effective for dimensionality reduction because it retains the majority of the original variance in fewer dimensions, minimizing information loss. This orthogonal transformation also eliminates multicollinearity, enhancing interpretability and efficiency in downstream analyses like clustering or regression. The method's focus on variance maximization ensures that the reduced dataset maintains critical patterns, making PCA a go-to method for exploratory data analysis and feature engineering [19].

Consider an $n \times p$ data matrix, X, with the sample mean of each column shifted to zero, where each of n rows represents a sample participant and each of the p columns represents a distinct feature. Mathematically, the transformation is defined by a set of size l of p-dimensional vectors of weights or coefficients $\mathbf{w}_{(k)} = (w_1, \dots, w_p)_{(k)}$ that map each row vector $\mathbf{x}_{(i)} = (x_1, \dots, x_p)_{(i)}$ of $\mathbf{X}$ to a new vector of principal component scores $\mathbf{t}_{(i)} = (t_1, \dots, t_l)_{(i)}$, given by

$$t_{k(i)} = \mathbf{x}_{(i)} \cdot \mathbf{w}_{(k)} \quad \text{for} \quad i = 1, \dots, n \quad \text{and} \quad k = 1, \dots, l$$

in such a way that the individual variables $t_1, \dots, t_l$ of $\mathbf{t}$ considered over the data set successively inherit the maximum possible variance from $\mathbf{X}$, with each coefficient vector $\mathbf{w}$ constrained to be a *unit vector* (where l is usually selected to be strictly less than p to reduce dimensionality).

The above may equivalently be written in matrix form as

$$\mathbf{T} = \mathbf{X}\mathbf{W}$$

where $\mathbf{T}_{ik} = t_{k(i)}$, $\mathbf{X}_{ij} = x_{j(i)}$, and $\mathbf{W}_{jk} = w_{j(k)}$.

First component: In order to maximize variance, the first weight vector $\mathbf{w}_{(1)}$ thus has to satisfy,

$$\mathbf{w}_{(1)} = \arg\max_{\|\mathbf{w}\|=1} \left\{ \sum_i (t_1)^2_{(i)} \right\} = \arg\max_{\|\mathbf{w}\|=1} \left\{ \sum_i \left(\mathbf{x}_{(i)} \cdot \mathbf{w}\right)^2 \right\}$$

3.2 Evaluating the impact of CMV seropositivity using propensity score methods

Propensity scores and inverse probability weighting (IPW) are statistical techniques designed to mitigate confounding in observational studies by approximating the conditions of randomized controlled trials. The propensity score is defined as the conditional probability of receiving a treatment given pre-treatment covariates:

$$e(X) = Pr(Z = 1|X) = E(Z|X)$$

where Z indicates treatment assignment (1 = treated, 0 = control) and X represents observed covariates. This score possesses a critical balancing property, where $Z \perp X|e(X)$, meaning covariates become balanced between treatment groups conditional on the propensity score. Typically estimated via logistic regression, the propensity score transforms multidimensional covariate information into a single-dimensional value. Inverse probability weighting leverages these scores by assigning each subject a weight equal to the inverse of their probability of receiving the treatment they actually received: w = 1/P(treated|x) for treated individuals and w = 1/(1-P(treated|x)) for untreated individuals. These weights create a pseudo-population where the covariate distribution is independent of treatment assignment. The average treatment effect (ATE) is then estimated using the weighted means of outcomes:

$$ATE = \frac{\sum (Z_i \cdot Y_i/e(X_i)}{\sum Z_i/e(X_i)} - \frac{\sum (1 - Z_i) \cdot Yi/(1 - e(X_i))}{\sum (1 - Z_i)/(1 - e(X_i))}$$

By weighting observations inversely proportional to their treatment probability, IPW effectively corrects for selection bias, allowing for more accurate causal inference in non-randomized studies

3.3 Genes differentially expressed with CMV (DESeq)

Differential gene expression analysis (DESeq2) is a widely used computational tool for differential gene expression analysis in RNA-sequencing (RNA-seq) data, employing a negative binomial (NB) generalized linear model (GLM) to address the overdispersion characteristic of count data (variance > mean)[20]. Unlike the Poisson distribution, which assumes equal mean and variance, the NB model incorporates a dispersion parameter (α) to account for gene-specific variability, enabling robust handling of biological and technical noise. DESeq2 first normalizes raw counts using size factors to adjust for library depth differences, then estimates dispersion by sharing information across genes with similar expression levels—a process called shrinkage—to improve accuracy, particularly for low-count genes . The fitted NB model relates normalized counts (μ_{ij}) to experimental covariates via a log2-linear equation $\mu_{ij} = sizeFactor_j \cdot 2^{\beta_i X_j}$, where β_i represents log2 fold changes. Differential gene expressions are tested using Wald tests or likelihood ratio tests (LRT), generating p-values that are adjusted for multiple testing via the Benjamini-Hochberg method to control the false discovery rate (FDR). By integrating variance stabilization, gene-wise dispersion estimation, and rigorous statistical testing, DESeq2 balances sensitivity and specificity, making it a vital part of RNA-seq analysis.

DESeq2 automatically filters genes with low normalized mean counts during multiple testing correction and the threshold for gene filtering is empirically determined from the data to maximize the number of significant genes while controlling the false discovery rate (FDR). This filtering occurs in the results() step, removing genes with little power to detect differential expression.

3.4 Finding gene modules using WGCNA and Group Lasso

Weighted Gene Co-expression Network Analysis (WGCNA) is a systems biology method designed to identify clusters (modules) of highly correlated genes in large-scale transcriptomic datasets, enabling insights into functional relationships and associations with phenotypic traits[21]. The mathematical framework for module construction begins by calculating a pairwise correlation matrix between all gene pairs across samples. To emphasize strong correlations and suppress noise, a **soft-thresholding power** (β) is applied, transforming the correlation matrix into a weighted adjacency matrix ($a_{ij} = |\mathrm{cor}(g_i, g_j)|^\beta$)

$$\mathrm{TOM}_{ij} = \frac{\sum_u a_{iu} a_{uj} + a_{ij}}{\min\left(\sum_u a_{iu}, \sum_u a_{uj}\right) + 1 - a_{ij}}$$

TOM reduces spurious connections by integrating indirect interactions, ensuring biologically meaningful modules. Hierarchical clustering is performed on the TOM dissimilarity (1 − TOM), and dynamic tree-cutting algorithms partition the dendrogram into modules, grouping genes with similar expression patterns. Each module is represented by a module eigengene (first principal component), which captures the dominant expression trend and facilitates trait-module correlation analysis. This approach balances sensitivity to subtle co-expression patterns with robustness to noise.

Group lasso is a regularization method that extends the standard lasso by enabling variable selection at the group level, making it ideal for scenarios where predictors naturally form clusters (e.g., genes within biological pathways). The sparse group lasso regularizer is an extension of the group lasso regularizer that also promotes parameter-wise sparsity. It is the combination of the group lasso penalty and the normal lasso penalty, the sparse group lasso penalty will yield a sparse set of groups and also a sparse set of covariates in each selected group[22]. Mathematically, it minimizes the objective function:

$$\arg = \min_{\beta_g \in \mathbb{R}^{d_g}} \frac{1}{n} \left\| \sum_{g \in \mathcal{G}} \mathbf{X}_g \beta_g - \mathbf{y} \right\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \sum_{g \in \mathcal{G}} \sqrt{d_g} \|\beta_g\|_2, \tag{1}$$

where $\mathbf{X}_g \in \mathbb{R}^{n \times d_g}$ is the data matrix corresponding to the covariates in group g, β_g is the regression coefficients corresponding to group g, $\mathbf{y} \in \mathbb{R}^n$ is the regression target, n is the number of measurements, d_g is the dimensionality of group g, λ_1 is the parameter-wise regularisation penalty, λ_2 is the group-wise regularisation penalty and $\mathcal{G}$ is the set of all groups.

The penalty term $\lambda_2 \sum_{g \in \mathcal{G}} \sqrt{d_g} \|\beta_g\|_2$ combines **L1 regularization over groups** (to induce group-wise sparsity) and **L2 regularization within groups** (to retain or exclude all variables in a group). By penalizing the L2 norm of each group's coefficients, the model shrinks entire groups to zero if their collective contribution is weak, preserving structural relationships within groups. This contrasts with the standard lasso, which indiscriminately shrinks individual coefficients. The $\sqrt{d_g}$ term adjusts for group size, ensuring fair penalization across heterogeneous groups. The method's convexity guarantees convergence, while its invariance to group-wise orthogonal transformations enhances robustness in correlated settings. Group lasso is particularly effective in genomic studies, survey analysis, and settings requiring hierarchical or interpretable variable selection.

3.5 Latent Modeling using Neural Network

Deep neural networks (DNN) are hierarchical computational models composed of multiple interconnected layers that transform input data through a series of non-linear mappings to learn complex patterns[23, 24]. Mathematically, each layer l in a network with L layers computes $h^{(l)} = \sigma(W^{(l)}h^{(l-1)} + b^{(l)})$, where $W^{(l)}$ represents weights, $b^{(l)}$ denotes biases, and σ is a non-linear activation function.

For binary classification tasks, BCE with logits loss efficiently combines the sigmoid function and binary cross entropy into a numerically stable formulation:

$$\mathcal{L}(y, \hat{y}) = -\frac{1}{n}\sum_{i=1}^{n}[y_i \cdot \log(\sigma(\hat{y}_i)) + (1 - y_i) \cdot \log(1 - \sigma(\hat{y}_i))]$$

where $\hat{y}$ represents raw logits and σ is the sigmoid function. This loss function avoids potential numerical issues when computing separate sigmoid and log operations.

For optimization, the Adam optimizer algorithm adapts learning rates for each parameter by estimating first and second moments of gradients: $m_t = \beta_1 mt - 1 + (1 - \beta_1)g_t$, $v_t = \beta_2 vt - 1 + (1 - \beta_2)g_t^2$, with bias-corrected updates $\hat{m}_t = \frac{m_t}{1-\beta_1^t}$, $\hat{v}_t = \frac{v_t}{1-\beta_2^t}$, resulting in the parameter update $\theta_t = \theta_{t-1} - \alpha \cdot \frac{\hat{m}_t}{\sqrt{\hat{v}_t}+\epsilon}$, where β_1 and β_2 are decay rates, g_t is the gradient of the objective function at timestep t, α is the learning rate, and ϵ is a small constant preventing division by zero. Adam's combination of momentum and adaptive learning rates makes it particularly effective for training deep neural networks with large datasets and complex loss landscapes[25].

4 Results

4.1 HRS Participant Characteristics

From Table 1, we observe that HRS participants have a median age of 69 years and a higher representation of Females (61.5%), Black (20.6%) and Hispanic (17.0%) individuals in the CMV seropositive group. Additionally, CMV seropositive participants exhibit elevated total inflammation scores and increased T cell counts compared to their seronegative counterparts.

Table 1: HRS Descriptive Statistics by CMV Status

	Negative (n=1081)	**Positive (n=2442)**	**p-value**
Biological Sex = F (%)	523 (48.4)	1503 (61.5)	<0.001
Racial Group (%)			<0.001
Non-Hispanic White	943 (87.2)	1439 (58.9)	
Non-Hispanic Black	72 (6.7)	503 (20.6)	
Hispanic	46 (4.3)	416 (17.0)	
Non-Hispanic Other	20 (1.9)	84 (3.4)	
Deceased (%)	95 (8.8)	303 (12.4)	0.002
Smoking Status (%)			0.021
Never smokers	482 (44.6)	1076 (44.1)	
Former smokers	503 (46.5)	1073 (43.9)	
Current smokers	96 (8.9)	293 (12.0)	
Age at 2016 (median [IQR])	68.00 [62.00, 76.00]	69.00 [62.00, 78.00]	<0.001
Education Years (median [IQR])	14.00 [12.00, 16.00]	12.00 [12.00, 15.00]	<0.001
Body Mass Index (median [IQR])	28.06 [24.80, 32.04]	28.13 [24.73, 32.25]	0.648
Inflammation Score (median [IQR])	-0.06 [-0.37, 0.28]	-0.01 [-0.33, 0.33]	0.005
# of comorbidities (median [IQR])	2.00 [1.00, 3.00]	2.00 [1.00, 3.00]	0.200
T cells count (median [IQR])	1.12 [0.83, 1.50]	1.36 [1.01, 1.79]	<0.001

4.2 Extracting TCR clones and clonal properties using MiXCR

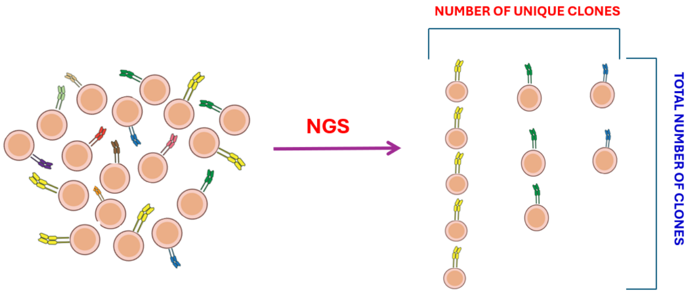


Figure 2: Clonal Properties

We processed paired-end FASTQ files from 4,388 samples, each containing complementary sequencing reads. Using MiXCR, we aligned these reads to the V, D, J, and C gene segments and successfully assembled complete TCR sequences for 3,757 participants. MiXCR jobs were run in batch using SLURM on the MSI Agate cluster. We then extracted the total number of clones and the number of unique clones for the α and β chains, excluding non-productive rearrangements. The clonal properties were highly correlated with each other and had a strong correlation with T-cell subsets (Figure 3).

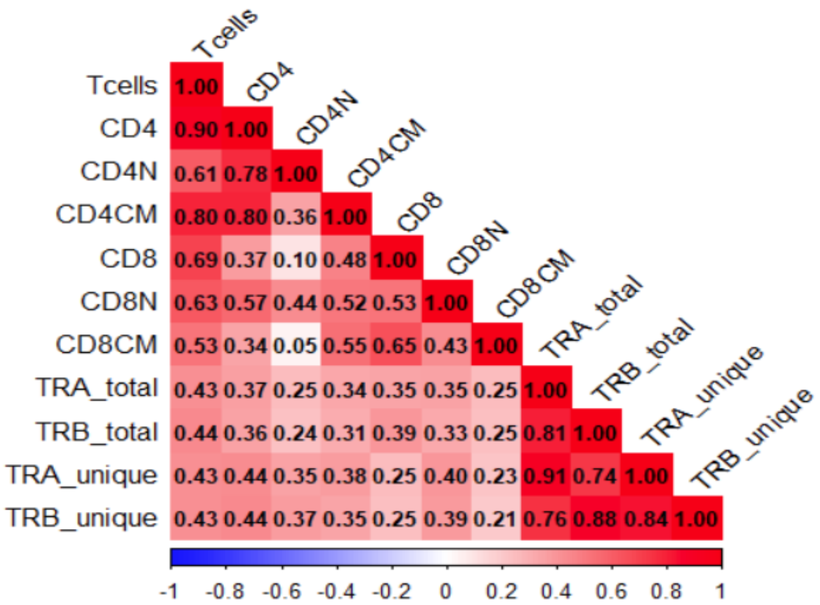


Figure 3: Correlation of TCR chains with T cell subsets

4.3 Association of TCR chains with all-cause mortality

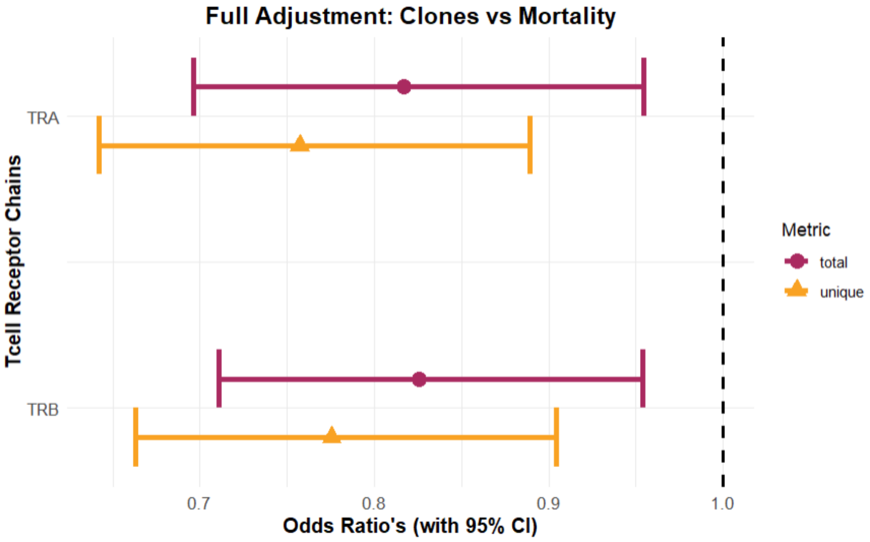


Figure 4: Clonal properties and mortality

We integrated clinical data from the HRS with clonotype properties extracted using MiXCR, resulting in a final dataset of 3,523 participants with valid CMV status, mortality information, and TCR chain metrics.

Then we evaluated the associations between total and unique clone counts of TCR-α (TRA) and TCR-β (TRB) chains and mortality in the HRS dataset comprising 3,523 participants. Models were adjusted for demographic factors, sequencing depth, smoking status, CMV seropositivity, BMI, inflammation, and comorbidity index. Higher TRA total (OR: 0.81; 95% CI: 0.70–0.95; p = 0.012) and unique clone counts (OR: 0.76; 95% CI: 0.64–0.89; p = 0.0008) were significantly associated with lower odds of mortality. Similarly, higher TRB total (OR: 0.83; 95% CI: 0.71–0.95; p = 0.01) and unique clone counts (OR: 0.77; 95% CI: 0.66–0.90; p = 0.0013) were also significantly associated with reduced mortality odds.

4.4 CMV and T-cell repertoire

Combining TRA and TRB clonal properties:

As illustrated in Figure 3, the clonal properties of the α-chain (TRA_total and TRA_unique) show strong correlations both within themselves and with the corresponding β-chain metrics (TRB_total and TRB_unique). To reduce dimensionality while preserving most of the variance, we performed Principal Component Analysis (PCA) on these four features. The first principal component (PC1), which captured approximately 86% of the total variance, was used as a composite measure of TCR clones (TRAB). The composite feature is then shifted by the min(TRAB) to have approximately similar scale to individual clonal properties.

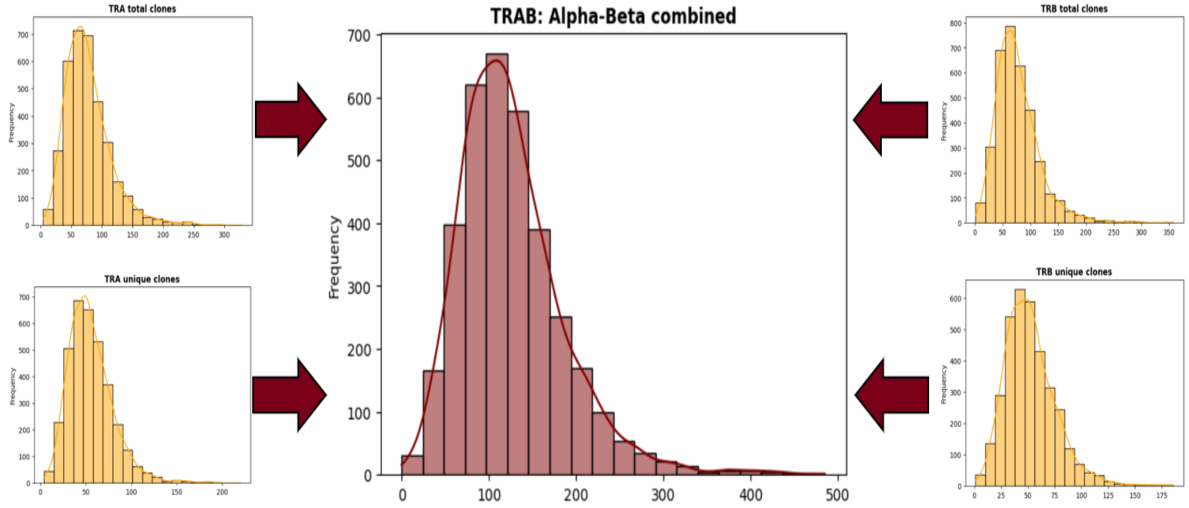


Figure 5: Combining clonal properties using PCA

Evaluating the impact of CMV seropositivity:

We weighted our sample of 3,523 participants using propensity score (PS) weights—PS stratified quintiles and inverse probability weighting—obtained from regressing CMV serostatus against age, sex, race, education years, smoking status, inflammation, comorbidity index, and total count of T-cells. We also ran an adjusted model with CMV and age interaction, and a model after one-on-one matching based on CMV status just for comparison. Standard errors were computed using 100 bootstrap samples with replacement.

The PS stratified models, commonly used in epidemiological studies, showed that CMV seropositivity leads to increased α/β (TRAB) clonal representation [ATE: 8.5, 95% CI: 4.5 to 12.5] and decreased percentage of CD4N [ATE: 6.6, 95% CI: -8.1 to -5.2] in comparison to seronegativity.

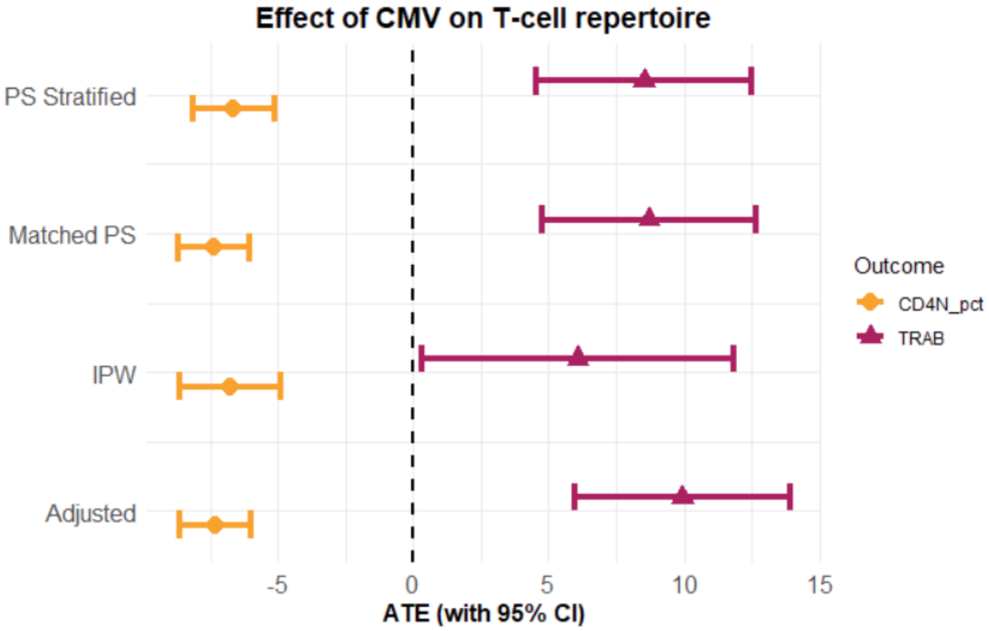


Figure 6: ATE of CMV seropositivity

Figures 4 and 6 suggest that CMV infection promotes clonal expansion of α/β TCR chains, which are each associated with reduced mortality. At the same time, CMV appears to deplete CD4 naive T-cells—a subset previously linked to lower mortality hazards—highlighting its complex and potentially opposing effects on immune aging. This is a paradox, and we looked into this by examining gene expressions that mediate the relationship between CMV infection and T-cell repertoire (Figure 7).

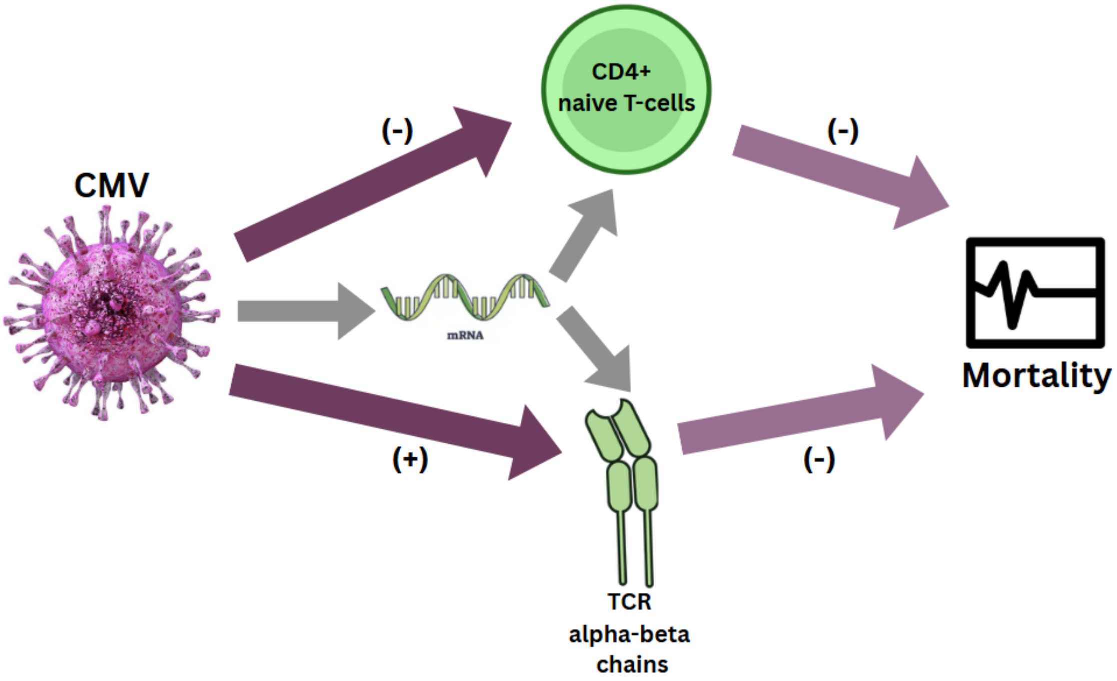


Figure 7: Analysis Flow Diagram

Gene Differentially Expressed with CMV:

We began with the HRS gene expression dataset comprising of 58,219 transcripts from 3,523 participants. To reduce noise and focus on informative genes, we first filtered for those with a mean expression level greater than 3 counts per million, yielding 12,800 genes. We then applied the DESeq2 pipeline to identify genes that were significantly differentially expressed between CMV groups, adjusting for age, sex, race, education level, smoking status, lymphocyte count, BMI, comorbidity index, and inflammation. After Benjamini-Hochberg correction (adjusted p-value < 0.05), **6,202 genes** remained as significantly differentially expressed.

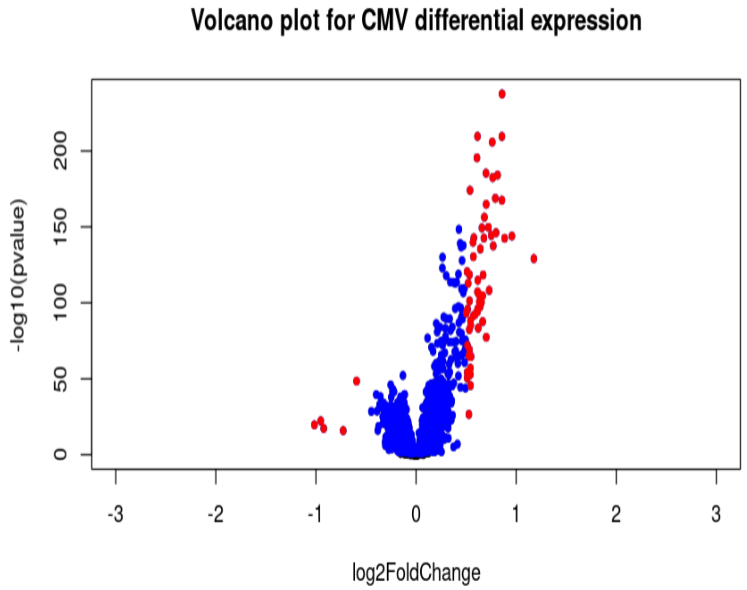


Figure 8: DESeq2 Volcano Plot

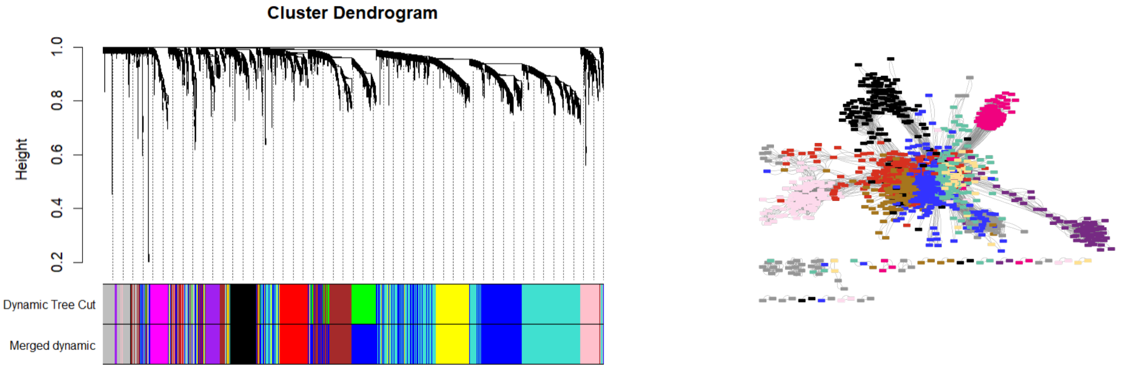
From the volcano plot, we can see that there are more up-regulated genes than down-regulated genes. This suggests that CMV exposure is associated with a broad increase in gene activity across the genome.

Finding Co-Expression Modules:
Starting with the differentially expressed genes associated with CMV from 3,523 participants, we performed hierarchical clustering using Euclidean distance on log2cpm transformed expression values of those genes to identify and remove outlier samples. After excluding genes with variance below the first quartile of all gene variances, we retained 4,651 genes from 3,435 participants for downstream analysis.

We then computed pairwise Pearson correlations and applied a soft-thresholding power of 8 (empirically chosen to approximate scale-free topology) to construct a dissimilarity matrix based on the Topological Overlap Matrix (TOM). Using these TOM-based dissimilarities, we identified gene co-expression modules. Finally, by merging modules whose eigengenes had a high correlation (i.e., dissimilarity < 0.2), we obtained 10 co-expression modules of varying sizes.

Table 2: Co-expression module sizes

black	**blue**	**brown**	**grey**	**magenta**
254	1139	500	413	164
pink	**purple**	**red**	**turquoise**	**yellow**
200	144	294	1124	419



(a) Gene modules with dendrograms

(b) Topological representation

Figure 9: WGCNA co-expression modules

After identifying the co-expression modules, we evaluated their predictive value for TCR clones (TRAB) and CD4 naïve T cells (CD4N) using separate Group Lasso models, applying a penalty exclusively at the group level. The optimal regularization parameter (λ_2) was selected based on the highest R^2 achieved via five-fold cross-validation. For TRAB, the best performance was observed at $\lambda_2 = 0.025$, resulting in six modules with non-zero coefficients. Similarly, for CD4N, $\lambda_2 = 0.01$ yielded seven informative modules. Four modules—blue (1139 genes), brown (500), black (254), and pink (200)—were shared between the two outcomes, suggesting a potential common regulatory signature.

Gene Ontologies and Pathway Enrichment:

In the case of the **blue module**, we observed enrichment for Gene Ontology (GO) terms related to taxis, chemotaxis, and immune signaling regulation. Additionally, KEGG pathway analysis revealed enrichment for pathways associated with cytomegalovirus (CMV) infection and cytokine-cytokine receptor interaction. Hence, we can deduce that this module is related to **T cell recruitment and activation** during a viral challenge.

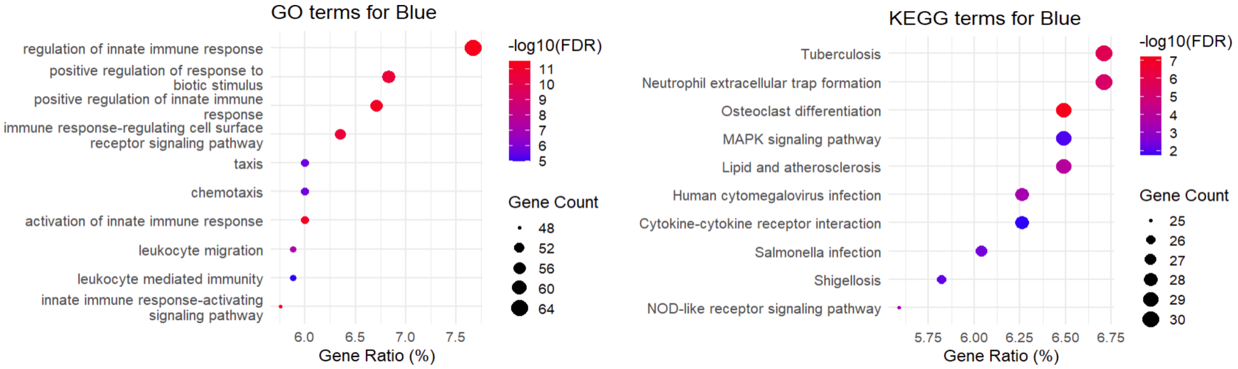


Figure 10: Blue module - enriched terms

For the **brown module**, Gene Ontology terms were enriched for processes related to the positive regulation of T cell responses, suggesting a role in **immune activation or potential dysregulation**, such as aberrant T cell expansion in autoimmune conditions. Notably, no KEGG pathways were enriched for this module.

The **black module** is enriched for GO terms related to **immune cell adhesion and activation**, and KEGG pathway enriched was for cytokine-cytokine receptor interaction. Together, these factors imply that the black module may orchestrate intercellular signaling events crucial for mounting an effective immune defense.

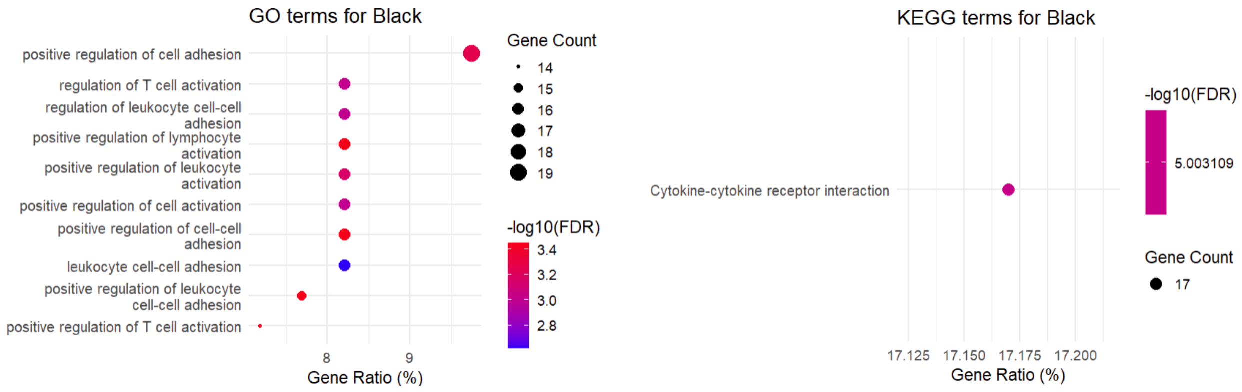


Figure 11: Black module - enriched terms

And the pink module is enriched for several immune-related GO terms, with the highest Gene Ratio (%) for components related to adaptive immune response and cell trafficking. Furthermore, enrichment for cell killing indicates potential involvement in cytotoxic immune mechanisms. KEGG pathways related to signalling were enriched for this module.

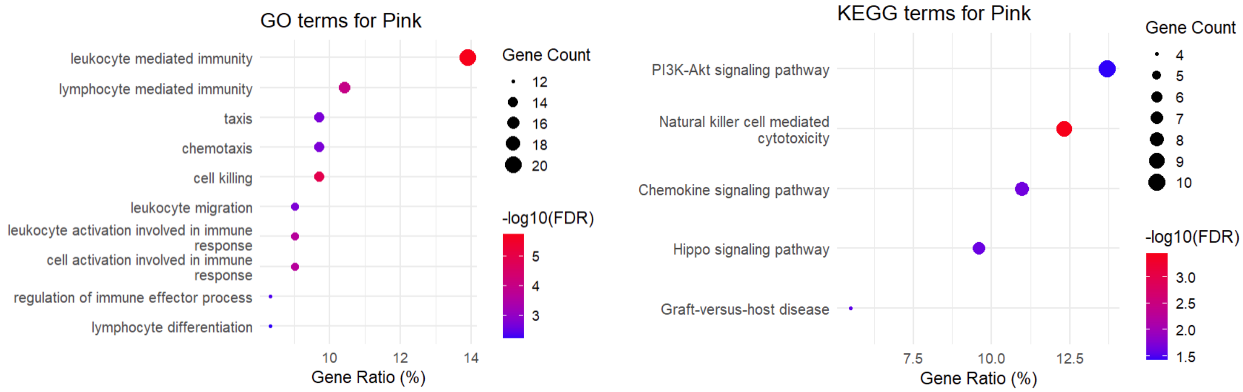


Figure 12: Pink module - enriched terms

Latent Modeling:

We developed a deep neural network composed of four parallel sub-networks, each corresponding to a specific module, to assess which of the underlying biological processes associated with these modules contribute to the reduced mortality risk linked to TRAB and CD4N.

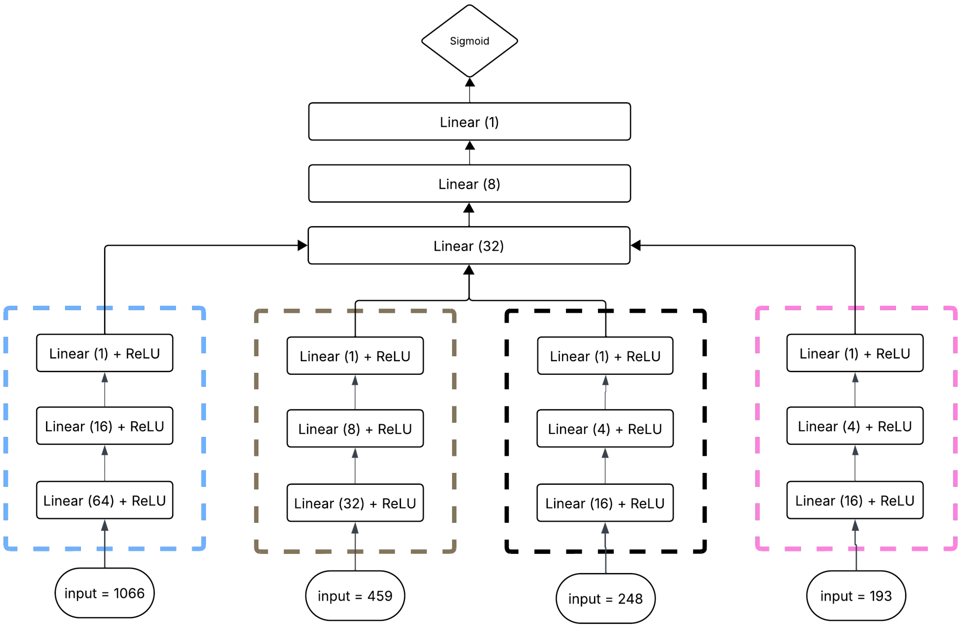
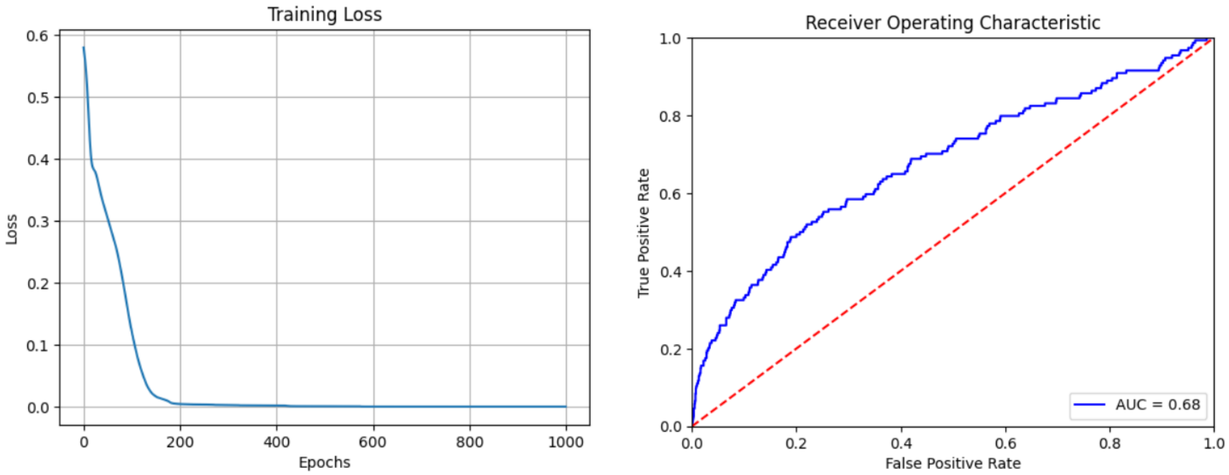


Figure 13: Deep Neural Network Architecture

Since we plan to validate our findings in the LLFS cohort, we selected only those genes within each module that overlap with the LLFS RNA-seq dataset.

In the HRS cohort, the model was trained for 1000 epochs on a training set of 2061 samples using a learning rate of 0.001. The model showed good convergence, achieving an AUC of 0.68 on the test set comprising 1374 samples.



(a) Training loss curve

(b) HRS Test-Set performance

Figure 14: Model results in HRS

In addition to producing predictions via the final sigmoid function, the neural network outputs latent representations of each module as 1D tensors. These representations are then used as inputs to a logistic regression model to assess the impact of each module—reflecting underlying biological pathways—on mortality after adjustment for age, sex, race, CMV, education, and BMI in the test set.

Table 3: Module Associations with Mortality in HRS Test-set

Predictors	**Odds Ratio**	**Lower CI**	**Upper CI**	**P-value**
blue	0.66	0.53	0.82	0.0002
brown	1.45	1.2	1.73	0.0001
black	0.82	0.65	1.04	0.097
pink	0.75	0.61	1.91	0.0048

Based on the logistic regression results on the test set (Table 3), only blue and pink modules showed a statistically significant association with reduced odds of mortality, where the predictors correspond to the latent representations of the gene modules obtained through the neural network trained on mortality.

We compared these latent representations with module eigengenes derived directly from WGCNA, which correspond to the first principal component of the genes within each module. Using the same architecture as in Figure 12, but substituting the sub-networks with module eigengenes, we achieved the test set performance of AUC = 0.64.

Table 4: Eigengene Associations with Mortality in HRS Test-set

Predictors	Odds Ratio	Lower CI	Upper CI	P-value
MEblue	0.94	0.72	1.21	0.63
MEbrown	1.07	0.81	1.41	0.62
MEblack	0.72	0.54	0.95	0.02
MEpink	0.83	0.63	1.09	0.17

Logistic regression results (Table 4) shows that module eigengene of blue and pink modules are not significant in the test set. That indicates that our latent model—which captures non-linear interactions within module and between modules—is a better way to capture module representations than module eigengenes.

We also evaluated whether the modules acts as surrogate for the CD4N/TRAB (results nor shown). We used multivariate logistic regression model to find the odds ratios for CD4N/TRAB against mortality while accounting for the module representation. We found that including all the modules together in the model in-addition to covariates make the impact of CD4N close to negligible and TRAB to lose its statistical significance on mortality.

Validation in LLFS:

The LLFS data consists of participants with RNA-seq measurements from both probands and offspring generation along with their spouses in visit-1 (N=1302). For more information about participant characteristics of this cohort refer to **Table 5**.

Table 5: LLFS Descriptive Statistics by CMV Status

	Negative (n=553)	Positive (n=749)	p-value
Biological Sex = F (%)	278 (50.3)	439 (58.6)	0.003
Generation (%)			<0.001
proband	92 (16.6)	246 (32.8)	
pro-spouses	17 (3.1)	38 (5.1)	
offspring	326 (59.0)	332 (44.3)	
off-spouses	118 (21.3)	133 (17.8)	
Deceased (%)	101 (18.3)	155 (20.7)	0.308
Age at visit-1 (median [IQR])	62.00 [55.00, 72.00]	70.00 [60.00, 88.00]	<0.001
Education Years (median [IQR])	13.00 [10.00, 14.00]	10.00 [9.00, 14.00]	<0.001
Body Mass Index (median [IQR])	26.70 [24.10, 29.75]	26.75 [23.90, 30.10]	0.931
C-Reactive Protein (median [IQR])	1.44 [0.76, 3.40]	1.72 [0.84, 3.70]	0.112
Interleukin-6 (median [IQR])	0.82 [0.49, 1.68]	1.15 [0.67, 2.10]	<0.001
T cells count (median [IQR])	1.23 [0.92, 1.58]	1.31 [0.99, 1.66]	0.015

In the LLFS cohort, the CMV positive group included a higher proportion of female participants (58.6%) and was, on average, eight years older than the CMV negative group. Additionally, CMV positive individuals exhibited elevated levels of inflammatory biomarkers (C-reactive protein and interleukin-6) as well as higher total T cell counts.

The neural network trained on HRS was applied to the LLFS cohort to extract latent representations for each gene module. These representations were then used in a logistic regression model to evaluate the association between the gene modules and mortality after adjustment for age, sex, field center, CMV, education, and BMI in the LLFS cohort (Table 6).

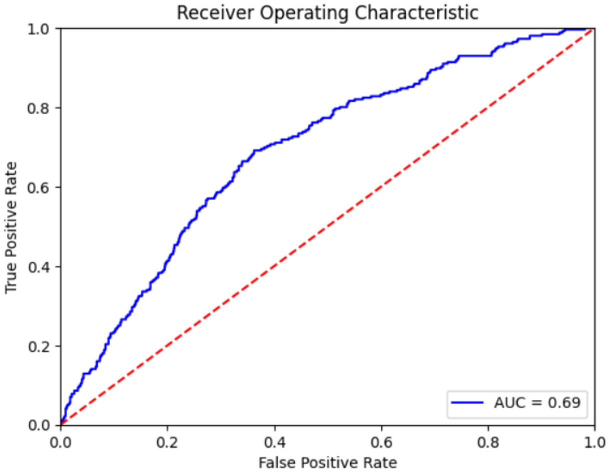


Figure 15: Model performance in LLFS cohort

Table 6: Module Associations with Mortality in the LLFS cohort

Predictors	Odds Ratio	Lower CI	Upper CI	P-value
blue	0.79	0.63	0.99	0.04
brown	1.3	1.06	1.59	0.01
black	0.83	0.62	1.11	0.22
pink	0.78	0.61	0.96	0.016

Consistent with the HRS findings, blue and pink modules remained significantly associated with reduced mortality odds in the LLFS cohort.

5 Discussion

Immune profiling of bulk RNA-seq data has been there for almost a decade, but **this is the first time that T-cell receptor clones have been extracted in a large multiethnic cohort**. The HRS had bulk RNA-seq data from 2016, from which we extracted

TCR sequences using MiXCR, and then calculated the total and unique clone counts for both TCR-α (TRA) and TCR-β (TRB) chains. We found that each of these measures was independently associated with reduced odds of mortality (Figure 4), even after adjusting for CMV serostatus, systemic inflammation, and comorbidity index. Given the high correlation among the TCR clone metrics, we derived a latent variable to represent their overall diversity, TCR α-β (TRAB) (Figure 5). Traditionally, T-cell repertoire assessments focus solely on TCR diversity. However, this overlooks an important dimension of T-cell immunity: versatility. To capture this, we also incorporated CD4+ naïve T cells (CD4N) as an additional component of the T-cell repertoire. CD4N has previously been shown to be associated with lower mortality risk.

A central question we aimed to address was whether a common chronic viral infection such as Cytomegalovirus (CMV) influences the T-cell repertoire. Our analysis revealed that CMV infection is associated with increased clonal expansion of TCR chains, yet it exerts a detrimental effect on CD4N levels (Figure 6). This creates an intriguing paradox: CMV drives opposing effects on two components of the T-cell repertoire, despite the fact that each of them individually was protective against mortality (Figure 7). A major public health implication of this paradox is that it could be a sign of "beneficial exhaustion", where expanded memory/effector clones being functional are strong enough to handle most threats encountered during aging. And, identifying the pathways underlying this beneficial exhaustion could offer promising targets for future immunological interventions

Our analysis of co-expression modules derived from genes differentially expressed with CMV indicates that CMV influences the T-cell repertoire through distinct biological processes. These include: cell trafficking and immune signaling regulation (blue module), enriched for CMV-related and cytokine-cytokine receptor interaction pathways; positive regulation of T-cell responses (brown module); immune cell adhesion and activation (black module), also linked to cytokine signaling; and immune response and cell trafficking (pink module), enriched for chemokine signaling pathways. Among these, the blue module showed the most robust enrichment in terms of gene count and KEGG pathways directly related to CMV, followed by the black module, which also shared enrichment in cytokine signaling. These findings suggest that CMV's contrasting effects on different components of the T-cell repertoire are largely mediated by biological processes involving cell trafficking, adhesion, and activation—driven primarily through cytokine signaling mechanisms.

A novel aspect of our analysis is the use of a deep neural network with dedicated sub-networks to derive module representatives, instead of relying on WGCNA-derived eigengenes. This approach more effectively captured non-linear interactions within modules while also accounting for between-module relationships in assessing mortality risk.

Notably, among the identified pathways, cell trafficking related modules (blue and pink) emerged as key contributors to the protective association between the T-cell repertoire and reduced mortality. This finding was further validated in the LLFS cohort, where the association remained significant. Another notable finding is that the brown module was associated with increased odds of mortality in both cohorts. This module is enriched for pathways involved in the positive regulation of T-cell responses—a process typically beneficial but potentially detrimental in the context of autoimmunity. This paradoxical association warrants further investigation.

Overall, our analysis identified that cell trafficking pathway was strongly associated with reduced mortality risk, while also being significantly impacted by CMV. Since our findings apply to both a general population of older adults (HRS) and a unique population with familial longevity (LLFS), they highlight a potential target for immunological interventions aimed at mitigating the adverse effects of chronic viral exposure on the T-cell repertoire. In future work, we plan to leverage longitudinal data on T-cell repertoire components to gain a more robust understanding of CMV's impact over time. Additionally, we will employ a gene network approach that allows for overlapping clusters, enabling a more accurate approximation of gene interactions and regulatory mechanisms. Although the findings of this study were validated in an external cohort, several limitations should be acknowledged. The primary analyses were conducted in the HRS, which includes participants aged 55 and older. As such, the observed associations may partly reflect age-related epigenetic modifications, limiting the generalizability of these findings to young populations. Additionally, our latent variable TRAB, derived from the α and β TCR chains in bulk RNA-seq data, provides only a coarse-grained view of T-cell diversity and function, lacking the resolution offered by single-cell approaches. Nevertheless, this study represents a pilot effort in exploring the broader impact of chronic viral infections like CMV on the T-cell repertoire.

6 Conclusion

This is the first study to investigate the impact of chronic viral infection, specifically Cytomegalovirus (CMV), on the T-cell repertoire, including CD4+ naïve T cells. Our findings demonstrate that while CMV infection drives the clonal expansion of TCR clones, it exerts a detrimental effect on the CD4+ naïve T cell population. These opposing impacts appear to be mediated through biological processes related to cell trafficking, adhesion, and immune activation—primarily orchestrated by cytokine/Chemokine signaling. Notably, among these pathways, cell trafficking emerged as a key contributor to the observed inverse association between the T-cell repertoire and reduced mortality. This relationship was further validated in an independent external cohort.

Code Availability

The code for our analysis on HRS and LLFS datasets are present in Git repository.

Acknowledgments

I would like to thank the following:

- **Dr. Bharat Thyagarajan**, for providing access to the datasets and for his guidance throughout the project.
- **Dr. Ryan Martinez**, for sharing his immunological expertise.
- **Faculty members of ARDL**, for their ideas and support in refining this work.
- **Committee members**: Dr. Adam Rothman, Dr. Bharat Thyagarajan, Dr. Weihua Guan, Dr. Aaron Molstad, and Dr. Erich Kummerfeld, for their support and valuable feedback.
- **Minnesota Supercomputing Institute (MSI)**, for providing their computing resources.
- The participants and data collectors of the HRS and LLFS studies, for their invaluable contributions.

I also wish to thank my family for their endless support during turbulent times.

References

1. Castelo-Branco, C. & Soveral, I. The immune system and aging: a review. en. *Gynecological Endocrinology* **30,** 16–22. ISSN: 0951-3590, 1473-0766. https://www.tandfonline.com/doi/full/10.3109/09513590.2013.852531 (2023) (Jan. 2014).
2. Mahdy, A. K. *et al.* Bulk T cell repertoire sequencing (TCR-Seq) is a powerful technology for understanding inflammation-mediated diseases. en. *Journal of Autoimmunity* **149,** 103337. ISSN: 08968411. https://linkinghub.elsevier.com/retrieve/pii/S0896841124001719 (2025) (Dec. 2024).
3. Dessalles, R. *et al.* How Naive T-Cell Clone Counts Are Shaped By Heterogeneous Thymic Output and Homeostatic Proliferation. *Frontiers in Immunology* **12,** 735135. ISSN: 1664-3224. https://www.frontiersin.org/articles/10.3389/fimmu.2021.735135/full (2025) (Feb. 2022).
4. Gaimann, M. U., Nguyen, M., Desponds, J. & Mayer, A. Early life imprints the hierarchy of T cell clone sizes. eng. *eLife* **9,** e61639. ISSN: 2050-084X (Dec. 2020).
5. Emerson, R. O. *et al.* Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. en. *Nature Genetics* **49,** 659–665. ISSN: 1061-4036, 1546-1718. https://www.nature.com/articles/ng.3822 (2025) (May 2017).
6. Russell, M. L. *et al.* Combining genotypes and T cell receptor distributions to infer genetic loci determining V(D)J recombination probabilities. en. *eLife* **11,** e73475. ISSN: 2050-084X. https://elifesciences.org/articles/73475 (2025) (Mar. 2022).
7. Seshadri, G. *et al.* Immune cells are associated with mortality: the Health and Retirement Study. *Frontiers in Immunology* **14,** 1280144. ISSN: 1664-3224. https://www.frontiersin.org/articles/10.3389/fimmu.2023.1280144/full (2023) (Oct. 2023).
8. Ramasubramanian, R. *et al.* Evaluation of T-cell aging-related immune phenotypes in the context of biological aging and multimorbidity in the Health and Retirement Study. en. *Immunity & Ageing* **19,** 33. ISSN: 1742-4933. https://immunityageing.biomedcentral.com/articles/10.1186/s12979-022-00290-z (2023) (Dec. 2022).
9. Huang, X. *et al.* Association between cytomegalovirus seropositivity and all-cause mortality: An original cohort study. en. *Journal of Medical Virology* **96,** e29444. ISSN: 0146-6615, 1096-9071. https://onlinelibrary.wiley.com/doi/10.1002/jmv.29444 (2025) (Feb. 2024).

10. Pourgheysari, B. *et al.* The Cytomegalovirus-Specific CD4$^+$ T-Cell Response Expands with Age and Markedly Alters the CD4$^+$ T-Cell Repertoire. en. *Journal of Virology* **81,** 7759–7765. ISSN: 0022-538X, 1098-5514. https://journals.asm.org/doi/10.1128/JVI.01262-06 (2025) (July 2007).

11. Lindau, P. *et al.* Cytomegalovirus Exposure in the Elderly Does Not Reduce CD8 T Cell Repertoire Diversity. en. *The Journal of Immunology* **202,** 476–483. ISSN: 0022-1767, 1550-6606. https://academic.oup.com/jimmunol/article/202/2/476/7953504 (2025) (Jan. 2019).

12. Lanfermeijer, J. *et al.* Age and CMV-Infection Jointly Affect the EBV-Specific CD8+ T-Cell Repertoire. *Frontiers in Aging* **2,** 665637. ISSN: 2673-6217. https://www.frontiersin.org/articles/10.3389/fragi.2021.665637/full (2025) (Apr. 2021).

13. Sidhom, J.-W. *et al.* Deep learning reveals predictive sequence concepts within immune repertoires to immunotherapy. en. *Science Advances* **8,** eabq5089. ISSN: 2375-2548. https://www.science.org/doi/10.1126/sciadv.abq5089 (2025) (Sept. 2022).

14. Gielis, S. *et al.* Detection of Enriched T Cell Epitope Specificity in Full T Cell Receptor Sequence Repertoires. *Frontiers in Immunology* **10,** 2820. ISSN: 1664-3224. https://www.frontiersin.org/article/10.3389/fimmu.2019.02820/full (2025) (Nov. 2019).

15. Sonnega, A. *et al.* Cohort Profile: the Health and Retirement Study (HRS). en. *International Journal of Epidemiology* **43,** 576–585. ISSN: 0300-5771, 1464-3685. https://academic.oup.com/ije/article-lookup/doi/10.1093/ije/dyu067 (2023) (Apr. 2014).

16. Barcelo, H., Faul, J., Crimmins, E. & Thyagarajan, B. A Practical Cryopreservation and Staining Protocol for Immunophenotyping in Population Studies. en. *Current Protocols in Cytometry* **84.** ISSN: 1934-9297, 1934-9300. https://onlinelibrary.wiley.com/doi/10.1002/cpcy.35 (2023) (Apr. 2018).

17. Wojczynski, M. K. *et al.* NIA Long Life Family Study: Objectives, Design, and Heritability of Cross-Sectional and Longitudinal Phenotypes. en. *The Journals of Gerontology: Series A* **77** (ed Le Couteur, D.) 717–727. ISSN: 1079-5006, 1758-535X. https://academic.oup.com/biomedgerontology/article/77/4/717/6420725 (2025) (Apr. 2022).

18. Bolotin, D. A. *et al.* MiXCR: software for comprehensive adaptive immunity profiling. en. *Nature Methods* **12,** 380–381. ISSN: 1548-7091, 1548-7105. https://www.nature.com/articles/nmeth.3364 (2025) (May 2015).

19. Zhang, Z. & Castelló, A. Principal components analysis in clinical studies. *Annals of Translational Medicine* **5,** 351–351. ISSN: 23055839, 23055847. `http://atm.amegroups.com/article/view/16014/16266` (2025) (Sept. 2017).

20. Anders, S. & Huber, W. Differential expression analysis for sequence count data. en. *Genome Biology* **11,** R106. ISSN: 1474-760X. `https://genomebiology.biomedcentral.com/articles/10.1186/gb-2010-11-10-r106` (2025) (Oct. 2010).

21. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. en. *BMC Bioinformatics* **9,** 559. ISSN: 1471-2105. `https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-559` (2025) (Dec. 2008).

22. Simon, N., Friedman, J., Hastie, T. & Tibshirani, R. A Sparse-Group Lasso. en. *Journal of Computational and Graphical Statistics* **22,** 231–245. ISSN: 1061-8600, 1537-2715. `https://www.tandfonline.com/doi/full/10.1080/10618600.2012.681250` (2025) (Jan. 2013).

23. Borisov, V. *et al.* Deep Neural Networks and Tabular Data: A Survey. *IEEE Transactions on Neural Networks and Learning Systems* **35,** 7499–7519. ISSN: 2162-237X, 2162-2388. `https://ieeexplore.ieee.org/document/9998482/` (2025) (June 2024).

24. Ryali, C., Nallamala, G., Fedus, W. & Prabhuzantye, Y. Efficient encoding using deep neural networks (2015).

25. Kingma, D. P. & Ba, J. *Adam: A Method for Stochastic Optimization* Version Number: 9. 2014. `https://arxiv.org/abs/1412.6980` (2025).