Associating Single-Cell Latent Factors with Genetic Risk

An analysis on a clinical patient cohort

Antonios Tsoukas[1]
Supervisor(s): Marcel Reinders[1]**, Inez den Hond**[1]**, Kirti Biharie**[1]
[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 21, 2026

Name of the student: Antonios Tsoukas
Final project course: CSE3000 Research Project
Thesis committee: Marcel Reinders, Inez den Hond, Kirti Biharie, Christoph Lofi

An electronic version of this thesis is available at http://repository.tudelft.nl/.

Abstract

Rheumatoid arthritis (RA) is a highly heritable disease, yet how its genetic risk translates into cell-type-specific mechanisms remains poorly understood. LIVI is a model that decomposes single-cell expression into donor and cell-state latent spaces, allowing for the reconstruction of the original data, but additionally leaving room for analysis of the retained latent information. The model has been shown to recover polygenic risk signals in healthy cohorts, but whether that is transferable to cohorts with active disease has not been tested. In this work, we apply LIVI to a predominantly RA cohort, with osteoarthritis (OA) patients as control, and ask whether the latent factors carry information about polygenic risk. After first confirming that the clinical cohort cell-state space recovers known immune cell populations, we test each of the 700 donor factors against the polygenic risk scores (PRSs) for 21 diseases, under different testing conditions, and find that one significant factor (D_{462}) is recovered between the latent space and RA PRS. This association survives ancestry correction, and changes in cohort. The factor localises to NK and T cells and drives antigen presentation program whose expression seems to be inversely related to RA risk.

1 Introduction and Background

Rheumatoid Arthritis (RA) is a chronic autoimmune disease that usually affects the joints in predominantly women of middle age or older, with an estimated genetic heritability of 60-70% [1]. It causes inflammation of the synovial joints and, if left untreated, can lead to joint deterioration and disability. Genome-wide association studies (GWAS) have been used to identify many genetic loci associated with risk for autoimmune genetic diseases like RA, but the molecular mechanisms behind these susceptibilities remain poorly understood [2]. These loci come in the form of single-nucleotide polymorphisms (SNPs), which appear as variants between individuals on the DNA. The largest genetic contributor to RA comes from the Human Leukocyte Antigen (HLA) region, although its contribution has been disputed recently, with older studies claiming 37%, while recent ones suggest the value is closer to 11% [1]. Resolving these mechanisms can have clinical implications, as identifying the genes and cell types through which genetic susceptibility is expressed could help identify patients early, and assist treatment.

Most of these SNPs found through GWAS however, fall outside of the protein coding genes, where the variant effects are self-explanatory, in regulatory DNA, where the production of genes is handled and the GWAS does not explain the causal mechanism [3]. Since a variant's effect is sometimes confined into a specific cell type or state, which is averaged away when expression is measured in bulk, single-cell RNA sequencing (scRNA-seq) is used. scRNA-seq assesses each cell individually, exposing gene expression at the desired resolution.

There are, however, two issues with scRNA-seq. Firstly, the vast number of variant-gene relationships make gene-level association testing intractable due to scale. Additionally, many of the effects of these variants are trans-effects, which act on genes anywhere on the genome, in contrast with cis-effects which affect their neighboring genes.

Recently, LIVI [4] addressed these issues by using a variational autoencoder to decompose scRNA-seq data into interpretable donor and cell-state latent factors that can be used for association testing with genetic risk scores. Since the latent space is learnt from expression rather than the genome itself, this association can be computed without the circularity of using the same genes to build and test the genetic score. In the LIVI paper, the model was validated on a cohort of *healthy* donors, where its donor factors appeared associated with polygenic risk signal for several autoimmune diseases, including RA. Whether the same

latent structure reflects genetic risk in an actual patient population, where risk is elevated and expressed as active disease, is still unknown.

This project addresses this gap with the following research question:

> *Do the latent factors of LIVI reflect genetic risk in a rheumatoid arthritis patient cohort?*

We investigate this using scRNAseq on a clinical cohort of 82 patients, 73 of which have been diagnosed with RA while the other 9 are Osteoarthritis (OA) patients. To do this we decompose the question into three sub-questions, addressed in order. We first ask whether LIVI's cell-state factors recover known immune cell populations when applied to our cohort. We view this as a prerequisite establishing that the latent space can meaningfully find the cell-states based on our scRNA-seq. Following that, we ask if LIVI's donor factors are significantly associated with RA polygenic risk scores across our cohort. Finally, for the associated factors, we inspect which cell states and genes drive the link to genetic risk, and whether they are consistent with the original LIVI findings. Recovering these would indicate that our results are a product of biology, and not a statistical artifact, strengthening our findings.

2 Related Works

2.1 Polygenic Risk Score

Polygenic Risk Score (PRS) is a metric that summarizes an individual's genome-wide genetic susceptibility for a trait or disease as a single value [5] . It is calculated as a weighted sum of SNPs, with weights given by the per-variant effect sizes estimated from the GWAS for the trait of interest [5, 6]. The number of variants included depends on the method used. Concretely, the score for individual i is

$$\mathrm{PRS}_i = \sum_{j=1}^{N} \beta_j \, x_{ij}, \tag{1}$$

where x_{ij} is the number of copies of SNP j in the genotype of individual i, β_j is that SNP's effect size from the GWAS, and N is the number of SNPs included in the score.

There are two major categories of methods for calculating PRS: Clumping and Thresholding (C+T), which removes variants before calculating the weighted sum, and genome-wide methods which include all variants while modelling linkage disequilibrium and applying shrinkage to the effect-size estimates [6]. In this case, C+T is used, as the removal of variants aids the prevention of redundancy, noise and overfitting by only keeping the strongest signals in each region of the genome (Section 5).

Importantly for this work, PRS provides us with a single, per-donor quantitative value which expresses the genetic susceptibility of the individual to a disease and can be used to test against other donor-level measurements from our data. There are, however, two important limitations of PRS that should be noted for this research. Primarily, PRS performance is very susceptible to variations in genetic ancestry, meaning that findings in one cohort might not transfer perfectly to another. An extension of this limitation is that predictive accuracy has been found to be consistently lower in non-European sample cohorts [5, 6]. To combat this, genetic principal components (PCs) are used in this work. These express the primary axes of genetic variation among the cohort, acting as a proxy for ancestry. By

using them as covariates, we remove the confounding effect of these ancestry differences, preventing population stratification from affecting any association. The second limitation is that individual-level estimates are inherently imprecise and can even differ between PRS calculation methods [6]. These limitations motivate us to interpret our results as a relation between a latent dimension and each donor's relative, population-calibrated genetic susceptibility, rather than an individual predictor.

2.2 Variational Autoencoders

The variational autoencoder (VAE) [7] is a generative model that learns a low-dimensional latent representation of high-dimensional data. It does this by coupling two jointly trained neural networks around a shared latent space, the former being called an encoder, with the latter referred to as the decoder. The former maps each observation x to a distribution over latent variables z, and the latter reconstructs the input based on samples of that distribution. The model is trained using amortized variational inference [8]. The latent space is what makes VAEs interesting for this work, as it turns intractable gene-level testing into association testing against a relatively small number of factors. We take advantage of this and try to find a relation to genetic risk through PRS.

2.3 LIVI model

LIVI, or Latent Interaction Variational Inference (Figure 1) [4], extends the VAE framework to decompose single-cell expression into three interpretable latent spaces, and to reconstruct them with a structured linear decoder. First of these spaces is the c latent space which captures cell-state information. The other two latent spaces D and V capture donor information, with the former expressing cell-state specific donor effects and the latter expressing global donor effects that are shared across all of a donor's cells, such as sex. Important to note for clarity is that these matrices are learnt by the model and are not provided as input. A linear decoder, decomposed to three decoders corresponding to each latent space (W_c, W_{DxC}, and W_V respectively) is used to reconstruct the original input from the latent spaces.

In operation, each input of the model comes in the form of a cell's gene expression vector $x \in \mathbb{R}^G$ from a donor. The encoder maps this cell to the the latent space and the donor ID can be used to map to rows of D and V. The decoders are used to reconstruct $\hat{x} \in \mathbb{R}^G$.

The cell-state-specific donor effects arise from a learnt assignment matrix $A \in \mathbb{R}^{15\times 700}$, which couples the 15 cell-state factors to the 700 donor factors via the interaction $cA \odot d_y$ (Figure 1), letting each donor factor act in a cell-state dependent way. Its columns map one-to-one onto D, so a D factor and its DxC factor (e.g. D_{462} / DxC_{462}) are interchangeable.

3 Results

The cohort used for this analysis contains 82 donors, with a total of 314,011 cells from synovial tissue. The donors are mostly female (73.2%), with a mean age of 57.9 (SD 14.8). Samples were taken from multiple synovial joints but mostly from the knee (48.8%). The amount of cells per donor also varies widely (median 2,834, IQR 964–6,452). These donors form the basis for all following association analyses (Table 1).

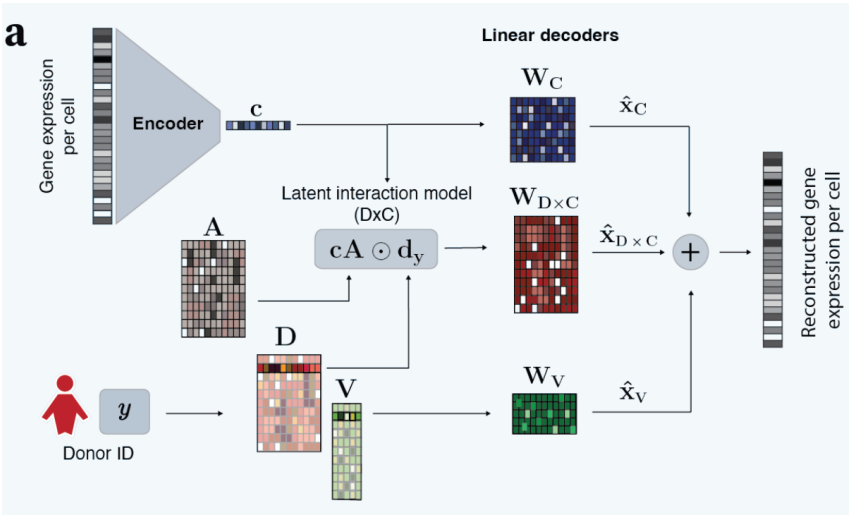

Figure 1: Overview of the LIVI model [4]. The latent spaces c, D and V can be seen with their corresponding decoders and the A matrix with the formula for moving from the donor to the cell state space.

Characteristic	Overall ($n = 82$)
Age, years - mean ± SD	57.9 ± 14.8
Sex, n (%)	
Female	60 (73.2)
Male	22 (26.8)
Tissue type, n (%)	
Knee	40 (48.8)
Wrist	31 (37.8)
MCP	5 (6.1)
Ankle	3 (3.7)
Hip	1 (1.2)
MTP	1 (1.2)
MTP 2	1 (1.2)
Cells per donor - median [IQR]	2,834 [964-6,452]

Table 1: Cohort characteristics ($n = 82$ donors, 314,011 cells).

3.1 The cell-state latent space recovers known immune populations

LIVI was trained with the scRNA data described above, learning the c, D and V latent spaces from them. PRS is a per-donor trait, meaning analysis has to happen on the donor level. It is, however, meaningful to first investigate the c space to inspect whether the decomposition of our data is faithful by retaining the information to distinguish between cell types. The c latent space was plotted using a UMAP (Figure 2).

The cell types form distinct separate clusters in the c space, coloured by the previously assigned cell-type annotations. Examples that stand out are the T cells and NK cells forming a joint cluster and B/plasma cells forming two similarly sized clusters. These can both be

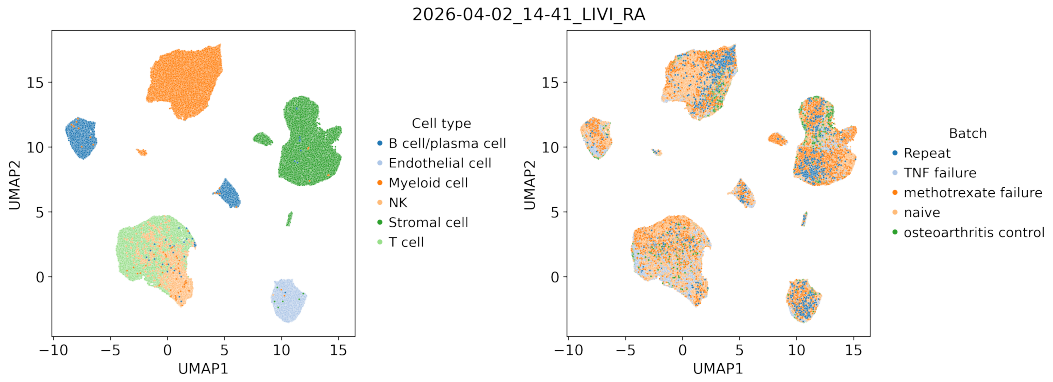

Figure 2: UMAPs of c space separation with labels of cell types (Left) and treatment batches (Right)

attributed to biology as NK and T cells are developmentally related, while B cells and plasma cells, which are categorized under the same label, are different states of a cell, which could cause the separation.

On the right labeled UMAP, we can see that there is no discernible separation or pattern of treatment batches, so treatment does not seem to structure the c latent space in a meaningful way. This analysis indicates that the cell-state space recovers cell-type information on clinical data, allowing us to proceed to test the donor factors of the D space, in order to look for an association with PRS.

3.2 Donor factors are associated with RA polygenic risk

The D latent space consists of 700-dimensional vectors, with each one relating to one donor. We test these donor factors for association with PRS, with the goal of locating factors that contain information related to the risk of the associated disease.

Each of the 700 donor factors of the D latent space was tested for association with the PRSs of 21 different diseases (Appendix A), which have been computed based on a GWAS by Ishigaki et al [9] (Section 5). Diseases include inflammatory arthritides such as rheumatoid, psoriatic, axial spondyloarthritis, autoimmune diseases such as Graves and Hashimoto, as well as multiple types of osteoarthritis and risk factor traits such as smoking and BMI. The association testing was done using Ordinary Least Squares (OLS) with age, sex and genetic principal components (PCs) as covariates, to mitigate confounding of population structure (Subsection 5.4). Across all of the tests, an association was only considered significant after the false discovery rate was corrected with the Benjamini-Hochberg procedure. The results presented here are the product of testing in the pooled cohort of all donors and the individual subsets of RA and OA.

After correction, a single donor factor was found to associate with RA PRS (D_{462}). Out of 700 factors tested against 21 PRS scores in all three cohort conditions, this factor is the only one to cross the significance threshold, visibly separated from the rest (Figure 3). The association is negative in both cohorts, with a β of -232.0 (adjusted $p = 0.024$) for the pooled cohort and β of -236.1 (adjusted $p = 0.017$) for the RA-isolated cohort (Table 2, Figure 4).

We note that, before the addition of the genetic principal components, the pooled cohort

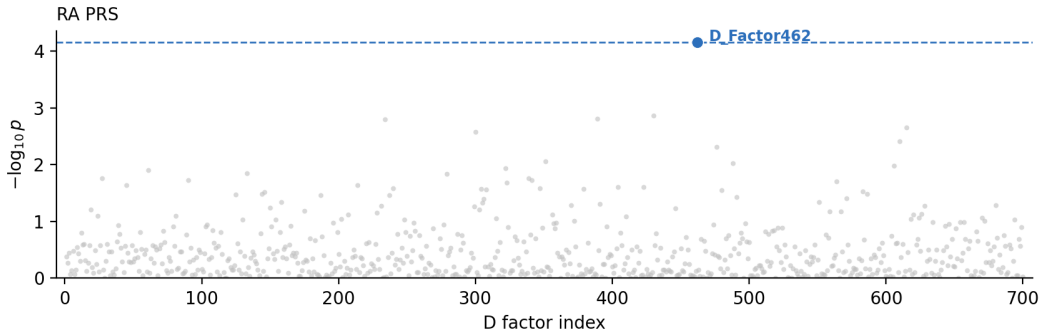

Figure 3: D factor association with RA PRS for the pooled cohort (The combined RA and OA cohort). The dashed line marks the Benjamini-Hochberg FDR threshold.

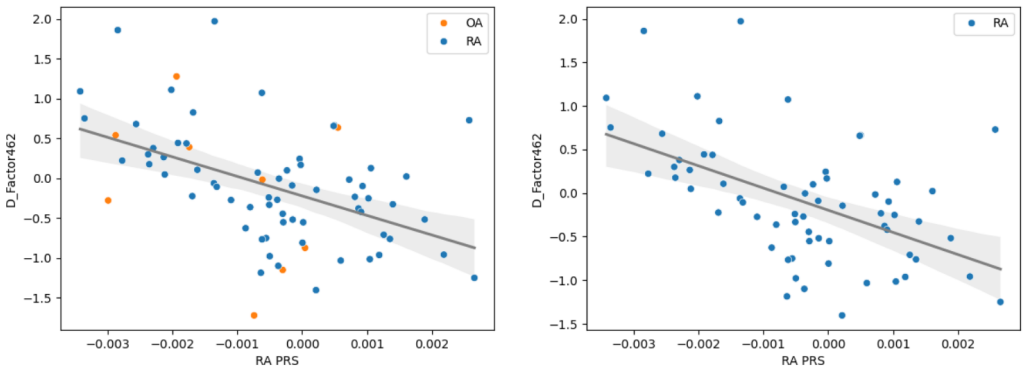

Figure 4: Scatterplots of significant D factor against RA PRS. Testing conditions from left to right: RA PRS in pooled cohort, RA PRS in RA cohort. Osteoarthritis donors are indicated in orange.

would additionally recover factor D_{329} when tested against the OA-finger PRS. This signal did not survive ancestry correction. We interpret this change as evidence that the inclusion of PCs removes signals that would otherwise be reported significant, and thus we do not report D_{329} in our findings (Appendix B). This leaves RA association as our only significant and robust result, which aligns with the cohort's consistency.

The size of the latent spaces is a hyperparameter picked in Vagiaki et al.[4]. However, no claims are made about the dimensions used being optimal, with the size of the D space specifically being proposed as an open avenue for future research in their discussion section. To investigate how these results scale with a smaller latent space, the exact same pipeline was run with a latent space from a LIVI model trained with 100 donor factors instead of 700. After correction, no significant factors were found under any of the experimental conditions.

3.3 The D latent space reflects non-HLA polygenic risk

The RA PRS used up until now excludes the HLA region. Since the region is the largest genetic contributor to RA [1], we wanted to investigate whether or not the association would be driven by HLA. We addressed this by computing an HLA-inclusive PRS and testing our

Factor	Condition	Effect size	p	p_{adj}
D_{462}	RA PRS, pooled	-232.253	0.000034	0.024046
D_{462}	RA PRS, RA cohort	-236.762	0.000024	0.017073

Table 2: Donor factors significantly associated with the RA PRS, across cohort runs. p_{adj} describes the p value after Benjamini Hochberg correction.

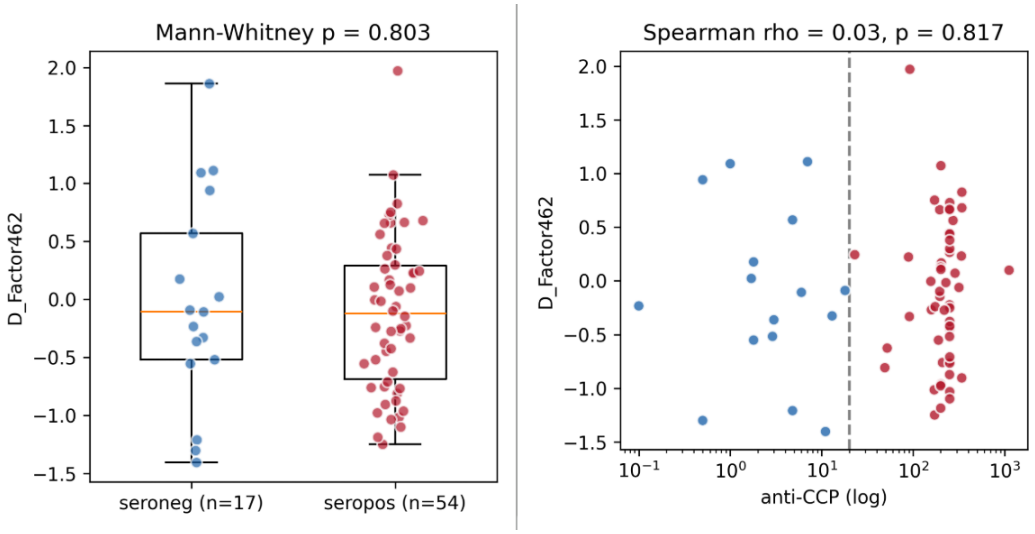

Figure 5: D_{462} is independent of serostatus. (a) D_{462} activity in seropositive ($n = 54$) and seronegative ($n = 17$) donors (ccp cutoff = 20). Medians are near-identical and groups do not significantly differ (Mann-Whitney $p = 0.80$). (b) D_{462} against ccp (log scale). No observed relationship (Spearman $\rho = 0.03$, $p = 0.82$). Donors colored by serostatus.

D factors against it.

The technique used to compute PRS before, C+T, has been shown to underperform on diseases with "large effects in regions of chromosome 6 with high LD" with autoimmune diseases being mentioned specifically [10]. Instead, we computed the scores as the sum of risk allele counts in the HLA region of the genome, weighted by their effect size for the disease. These PRS values were then tested against the D latent space in the same way as before, using OLS with the covariates of age, sex and genetic PCs. No D factor appeared significant from this testing, which was conducted against both the seropositive and seronegative RA PRS. D_{462} specifically was investigated, appearing insignificant against both PRSs ($\beta = -0.059$, adjusted $p = 0.953$ in seropositive and $\beta = 0.130$, adjusted $p = 0.997$ in seronegative case).

To further corroborate our results, we investigated if D_{462} separates the donors by serostatus. To annotate the donors we looked at ccp, a value describing the concentration of specific RA antibodies. We decided on a cutoff of 20 for our binary labels. After annotating the donors, the results of D_{462} between seropositive ($n = 54$) and seronegative ($n = 17$) donors appeared near identical (Mann-Whitney $p = 0.80$), and very similar medians. Additionally, a continuous test was conducted on ccp, which also concluded no relation between the factor and serostatus (Spearman $\rho = 0.03$, $p = 0.82$) (Figure 5).

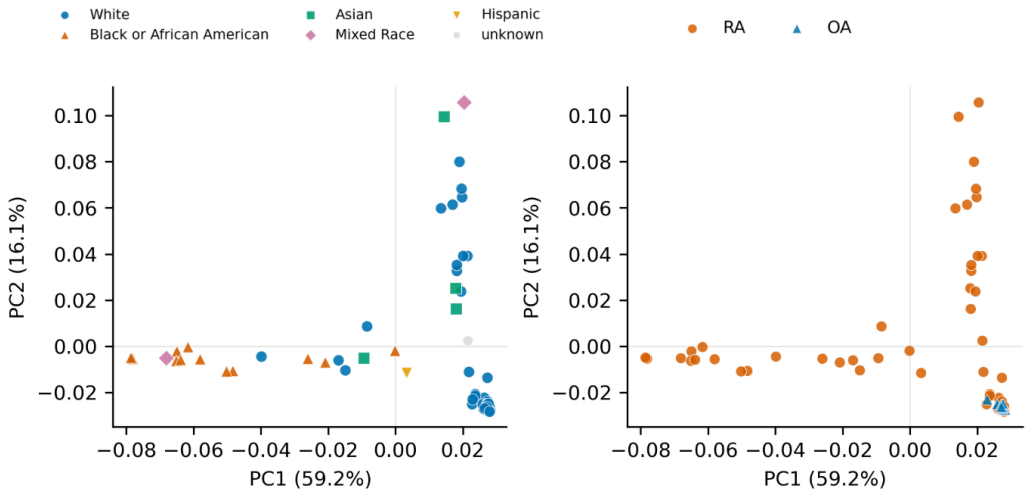

Figure 6: Scatterplots of first two genetic principal components (which explain approximately $75,3\%$ of the total variance). Additionally donors are annotated according to their race (Left) and disease (Right). The race acts as a proxy for ancestry. Expanded version of first 4 PCs can be found in Appendix C.

3.4 The Genetic Principal Components reflect ancestry

Since PCs are used as covariates on our research, it was considered to investigate them to see if they reflect actual ancestry, or some other cohort structure. This is done since the aim of using the PCs is removing ancestry confounding, so whether that is what they explain should be confirmed.

The PCs are computed based on the SNPs from the GWAS. Out of the ten primary PCs, the first four explain around $85,3\%$ of variance, with PC1 explaining the majority at $59,2\%$. After plotting and annotating the axes, they seem to track ancestry structure, with donors from the same populations creating clusters. Additionally, it can be noted that all OA samples seem to belong to the same ethnic group (Figure 6).

3.5 Cell types driving D462

To further explore the relationship of the latent space with RA genetic risk, we use D_{462} and investigate the cell types and genes connected to it and driving association. Using the cell-type $\times$ D matrix (Subsection 5.5), we observe that D_{462} places its largest weights on NK and T cells (Figure 7, Table 3). Importantly, since the association between the RA PRS and D_{462} is negative ($\beta = -232.0$ in the pooled cohort), this NK/T-localised activity is reduced in high risk donors. Softmax normalisation is applied to the c latent space during the construction of the matrix, but the version using the raw distribution is also presented in this work to display more of the variance of cell types and support the claims.

	D_{462} (RA)			
	mean		variance	
Cell type	raw	softmax	raw	softmax
NK cell	0.368	0.246	0.089037	0.000234
T cell	0.210	0.241	0.138412	0.000283
Endothelial cell	0.115	0.232	0.232764	0.000266
Myeloid cell	-0.125	0.201	0.164434	0.000579
B cell/plasma cell	-0.317	0.179	0.351751	0.000669
Stromal cell	-0.419	0.188	0.229748	0.000298

Table 3: Cell-type weight means and variance for the significant donor factor from the cell-type $\times D$ matrix, using the raw and softmax-normalised parametrisations. Rows are ordered by the raw D_{462} weight. It should be noted that variances are not comparable since they are computed within their respective spreads.

3.6 Genes and pathways driving the associated factors

To interpret the genes driving D_{462}, the decoder gene weights of the respective DxC factor (DxC_{462}) were used (Subsection 5.6). These were first inspected manually, by looking at the head and tail of the list, where the top individual loadings did not form a recognisable pathway, motivating us to use gene set enrichment analysis (GSEA) [11]. It is important to note that while the bottom loadings were equally uninformative, they did include genes like IGKC and IGLC3, which are expressed in B and plasma cells. This is consistent with our results in Subsection 3.5 where they appear as the lowest ranked cell type (Appendix D).

GSEA against the Reactome database [12] recovers antigen processing and presentation programs (ER-Phagosome, NES = 1.90, $q = 0.003$. Antigen processing-Cross Presentation, NES = 1.89, $q = 0.008$. Class I MHC peptide loading, NES = 1.88, $q = 0.006$). The leading genes in all of these comprise of HLA-C, B2M, TAPBP and HLA-B (Figure 8).

To connect this back to genetic risk, GSEA was additionally conducted against the genes linked to the SNPs used to compute our RA PRS. These genes are enriched near the top of the ranking (NES = 1.38, $q = 0.084$), passing the significance threshold of $q = 0.25$ (Subsection 5.6). The enrichment is largely HLA-driven with the top ranked gene being HLA-DRB1 (Figure 9). Excluding HLA genes lowers the peak enrichment score from approximately 0.59 to 0.5 and makes the results non-significant (NES = 1.18, $q = 0.253$). The leading genes are HLA-DRB1 and HLA-B, followed by non-HLA genes like PTPN2 and CTLA4, which remain at the top after the removal of HLA (Figure 9).

Both enrichments carry genes with positive D_{462} loadings (Figure 8, Figure 9), which fixes their direction relative to the factor. Based on the negative association from Subsection 3.2, we can interpret that, as the RA PRS is negatively related to the factor, and the factor is positively related to the antigen presentation program and relevant genes, the genes are inversely related to the genetic risk. This means donors with higher RA PRS are predicted to express the antigen-presentation program at lower levels in the cell states where D_{462} is active.

4 Discussion

This work asked whether the latent structure that LIVI learns from single-cell data continues to carry information related to genetic risk in a clinical cohort. Establishing this required

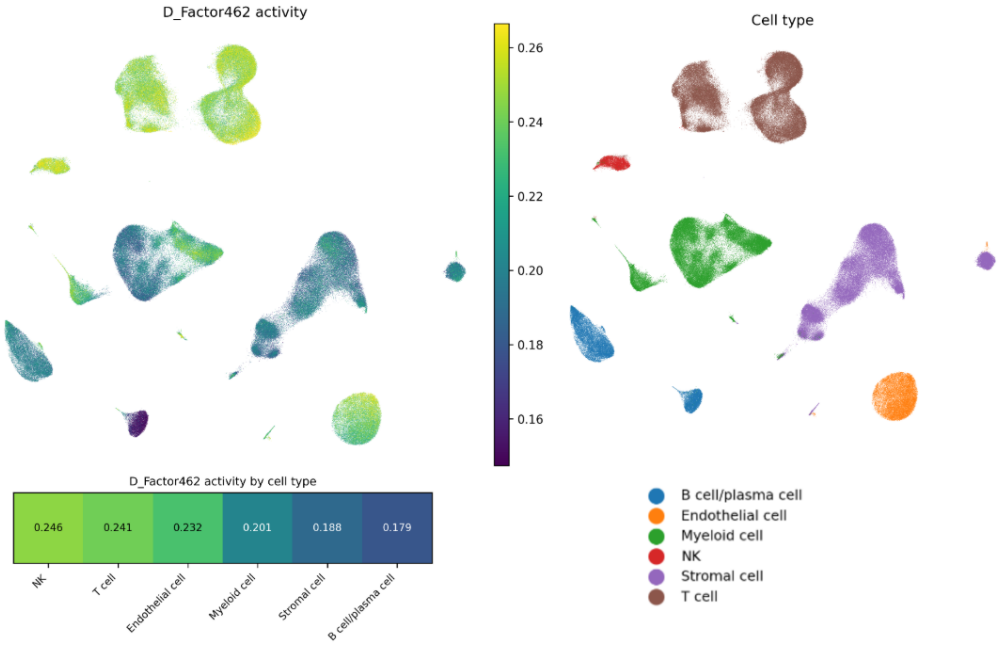

Figure 7: UMAP of cells colored by activity of D_{462} with softmax applied (Left) cell type (Right). Higher activity of the significant D factor can be seen in areas labeled as NK and T cells, which can be seen in the activity per cell type (Bottom)

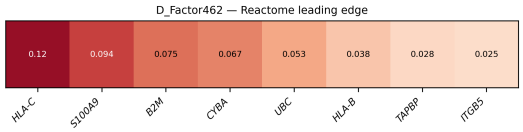

Figure 8: Leading-edge genes of Antigen processing-Cross Presentation (R-HSA-1236975 $NES = 1.89$, $q = 0.008$). Genes are ordered by loading.

three things: that the cell-state space recovers underlying biology, that some donor factors are significantly associated with a PRS value, and that those factors carry related biological meaning, rather than statistical artifacts. Our results satisfy all three, meaning we can affirmatively answer our research question of *Do the latent factors of LIVI reflect genetic risk in a rheumatoid arthritis patient cohort?*, with D_{462} carrying interpretable biological meaning related to RA PRS.

The first requirement is established by investigating the cell-state (c) latent space, where a clear separation of cell types can be observed, with no visible structure based on treatment batch (Figure 2). The cell type clusters that are formed corroborate known biology and confirm latent decomposition retains biological structure, allowing the following donor-level analysis to lean on a faithful representation rather than a technical structure.

The second requirement is met by D_{462}, which is found to have an association with RA PRS in both the exclusively RA cohort, along with the pooled cohort, which contains the additional OA donors. Three properties strengthen this finding, arguing towards it being a

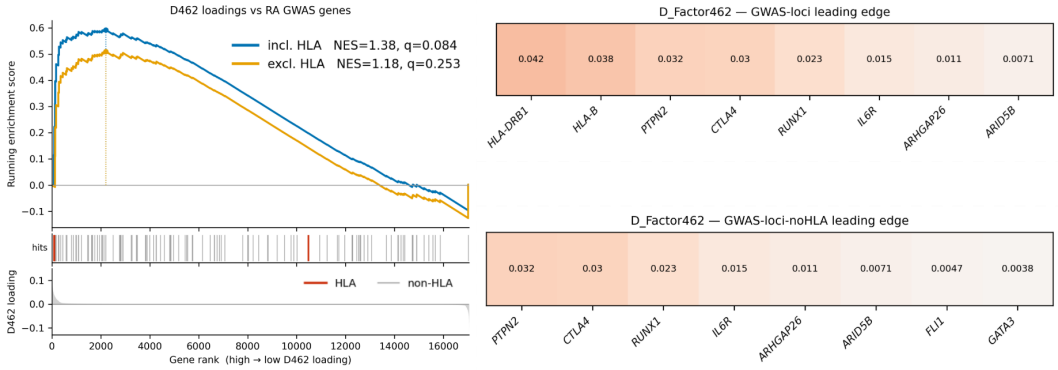

Figure 9: Preranked GSEA of the D_{462} loadings against the RA GWAS gene set. (Left) Running enrichment score including (blue) and excluding (orange) HLA genes. Peak ES and significance drop when HLA genes are removed (NES = 1.38, $q = 0.084$ - NES = 1.18, $q = 0.253$). Hit ticks colored by HLA membership. (Right) Leading-edge genes for the HLA-inclusive (top) and HLA-exclusive (bottom) analyses, ordered by loading.

genuine signal. First, it is uniquely specific, being the only factor out of 700 to cross the significance threshold when tested against 21 PRSs in three configurations. This, along with the fact that the hit is in the disease of our cohort, shows that the association between the latent space and genetic risk are not indiscriminate (Figure 3). Second, the signal is robust to ancestry, persisting after the addition of ten principal component covariates, which were confirmed to explain ancestry structure rather than other cohort features (Figure 6). This argues against the signal being a product of stratification. Third, the factor hit is reproduced across cohorts, with similar magnitude and direction in both the pooled and RA cohorts, while remaining absent from the OA cohort. This rules out the association being caused by noise from the inclusion of OA donors or some internal cohort composition.

The third requirement rests on the cell types and genes associated with D_{462}, and ties our findings back to biology. Mapping the factor to the cell state space through the cell-type $\times D$ matrix localises its activity to primarily NK and T cells (Figure 7). The factor is thus not evenly spread but focused on specific cell types. On the gene level, while initial investigation of the top loadings is not informative, enrichment against the Reactome database [12] recovers antigen processing and presentation programs, all led by the same genes (HLA-C, B2M, TAPBP and HLA-B, Figure 8). Testing the same ranked loadings against the genes of the RA GWAS shows that our risk loci genes sit near the top of D_{462}'s loadings (Figure 9). This is the most direct result of this work, as the genes D_{462} strongly expresses show overlap with the loci that defined our RA PRS, which ties the latent space to the disease's genetic architecture. An overlap can also be observed with the top loaded genes found by Vagiaki et al. [4] in their downstream analysis of their RA-related factor.

The association between D_{462} and RA PRS is negative ($\beta = -232.0$ in the pooled cohort), which warrants some interpretation. While the sign in a latent space is not an interpretable quantity on its own, its effect becomes apparent in downstream analysis. The quantity that does matter for this work is the product of this β and each individual gene loading. As D_{462} loads positively in antigen processing and presentation genes, the negative β implies an inverse relationship between RA genetic risk and the expression of these genes in NK and T cells. This direction is reported, but should be interpreted as descriptive and

not mechanistic. This is because our samples include genetic susceptibility, active disease and ongoing treatment, meaning that as the genetic risk is only one of these entangled axes, no concrete claim can be made about a causal relationship. What we can establish is that the association exists and targets the specific program, but not why the direction is set as it is.

The role of HLA in this analysis is worth some discussion. Neither the HLA-inclusive PRS nor donor serostatus showed any association with D_{462}, which indicates that the factor reflects non-HLA polygenic risk. As the PRS was computed as a weighted sum of HLA risk alleles rather than by clumping and thresholding, this null is meaningful and not an artifact of the scoring method. The genes captured through our interpretation include multiple HLA genes (HLA-C, HLA-B, HLA-DRB1), but these reflect the genes expressed by the factor, not the HLA risk alleles the donor carries. The risk captured by D_{462} therefore lies outside the HLA region.

Several limitations can be discussed on this work with the primary one involving the association model. In the current work, association is done using OLS, treating donors as independent, which is inaccurate. Including the covariates of age, sex and ten principal components attempts to remove these features from the equation, by addressing general genetic stratification and ancestry. They do not, however, address cross-donor relatedness, which covers the more granular genetic similarity. A linear mixed model (LMM) incorporating a kinship matrix would encode the covariance between donors, addressing the cross-relatedness directly. A second limitation involves the sample size of OA, which leads to a lack of statistical power. The absence of an associated factor cannot be read as interpretable evidence, and is better treated as uninformative. Third, PRS is imprecise on an individual level and is sensitive to ancestry, meaning the results should be read as a relationship between the latent space and a donor's relative, population-calibrated genetic risk, rather than an individual predictor. Lastly, the dependence on the latent dimensionality should be viewed as a limiting factor, as it is unclear how these results would scale, aside from our small experiment with the 100 factor model.

These limitations lead to multiple concrete future research paths. Importantly for replication, an investigation of if D_{462} or a corresponding factor is found in a separate independent clinical RA cohort would establish whether the captured biology is particular to this dataset. Replacing OLS with an LMM would remove relatedness confounding. Furthermore, using a model like a multilayer perceptron could allow multivariate analysis, testing whether genetic risk is spread among many factors rather than concentrated into one. A large OA cohort would allow the PRS to be tested properly instead of being underpowered. Finally, further research could be done as to whether D_{462} factor or the latent space at large can track treatment response of donors, which could create a link of clinical relevance.

5 Methodology

5.1 Data & Data Preprocessing

The data used for the model comes from the dataset of Zhang et al. [13]. It consists of samples from 82 individuals, out of which 73 are RA patients and 9 are control OA patients. 3 of the RA patients have repeat biopsies.

These biopsies were conducted by various different institutions and from various types of synovial tissue, including most joints like the wrist and knee, which are most prominent. The dataset presents a typical RA cohort, mostly consisting of women of middle age or

older, with a mean age of 57.9 and a standard deviation of 14.8. The cell count per donor has a median of $2,834$ (IQR $964 - 6,452$) which reflects heterogeneity in the amount of cells sampled from each donor. (Table 1).

For the calculation of the PRS a genome wide association study (GWAS) from Ishigaki et al. [9] was used. The PRS was then computed for our cohort based on these results (Subsection 5.3).

Importantly, the cohort used for most of the preprocessing, like calculating the PRS scores, is a superset of the Zhang et al cohort [13]. A side-effect of this is that some of the donors did not align, leading to a cohort of 73 samples, which are used for our analysis. It is important to note that all OA donors are included in the 73 samples, and the three donors with repeat biopsies are part of the subgroup of unused samples.

5.2 Experimental Setup

Our experimental setup begins by using a LIVI model trained on the aforementioned dataset to produce a latent space. The donor (D) factors are then tested for association with 21 PRS values for different diseases under three conditions: the pooled cohort, the RA donor samples, and the OA donor samples. The associated factors are then interpreted via gene loadings, cell-state contributions and enrichment.

5.3 Polygenic Risk Scores

For the calculation of the PRSs, a technique called "Clumping and Thresholding" ($C + T$) is used. The technique reduces the number of SNPs used by keeping the strongest signals in each region of the genome, using the p-values. The values are used to sort the SNPs based on importance. After that ordering, the top one is picked, and the SNPs that are in linkage disequilibrium (LD, which describes the correlation of nearby variants that tend to be inherited together) in a specified window around it are removed from the set (or "clumped" into it). This continues in a decreasing order of still available SNPs. The approach ensures that strong signals are retained, while keeping the selected SNPs independent, with low linkage disequilibrium. This is important as SNPs in LD arise from the same underlying signal, meaning including multiple would count the signal repeatedly, overrepresenting the region they came from. Restricting on independence ensures each SNP contributes new information to the score.

This approach was applied in order to create 21 different PRSs including for RA, OA-finger and multiple other OA types, as well as other diseases. These were then used for the rest of the analysis. The full list can be found in Appendix A

5.4 Association testing on the donor Level

To test for association, we regressed each factor against all 21 PRSs in the aforementioned cohorts using ordinary least squares (OLS). This analysis is univariate, as each factor is individually tested against the PRS, so the individual relation of each one can be investigated. To guard against confounding due to population stratification and ancestry, the ten first genetic principal components (PCs) were added as covariates, since differences in genetic ancestry have been shown to impact polygenic risk [5, 6]. The same was done with age and sex to additionally remove their effect from our testing. For donor i and factor d, our regression is expressed by the following formula:

$$y_{id} = \beta_0 + \beta_{\text{PRS}}\, s_i + \beta_{\text{age}}\, \text{age}_i + \beta_{\text{sex}}\, \text{sex}_i + \sum_{k=1}^{10} \beta_{PC,k}\, \text{PC}_{ik} + \varepsilon_{id}. \tag{2}$$

Due to the large number of test being ran, the p-values were corrected using the Benjamini-Hochberg procedure at a FDR of 0.05, with a factor being considered significant if the adjusted value is below this score. With each test of every PRS against the three cohorts, a table of results was created, containing for each factor the effect size β, the p-value and the adjusted p-value.

5.5 Interpretation of relevant cell types

The relevant cell types are found by creating a cell-type $\times$ D matrix as described in Vagiaki et al. [4]. The formula is the following:

$$\mathbf{M} \;=\; \mathbf{S}\,\tilde{\mathbf{C}}\,\mathbf{A}, \qquad \text{where } \tilde{\mathbf{C}} \text{ is } \mathbf{C} \text{ with softmax applied row-wise.} \tag{3}$$

Where $\mathbf{A} \in \mathbb{R}^{15\times 700}$ is the assignment matrix with 15 rows describing the c factors and 700 columns describing the DxC factors which directly correspond to D factors. $\tilde{\mathbf{C}} \in \mathbb{R}^{314011\times 15}$ is the row-wise softmax-normalised cell-state latent matrix over the 314011 cells, so each row is a distribution over the 15 cell-state factors summing to one. $\mathbf{S} \in \mathbb{R}^{6\times 314011}$ is the cell-type pooling matrix, with each row representing a cell type and holding uniform weights summing to one over all cells of the respective type. The result $\mathbf{M} \in \mathbb{R}^{6\times 700}$ has entries M_{td} equal to the mean activity of donor factor d across cells of type t.

The matrix can then be used to observe the breakdown of cell-types for a specific D factor, or to attain a 1:1 correspondence, mapping each D factor to a specific cell type, according to the highest weight.

For this work, these were additionally visualised using a UMAP of D_{462} activation across cells paired with a UMAP of cell types which can be cross-referenced. Additionally a single row visualisation has been added presenting the softmax-normalised averages of cell types (Subsection 3.5).

5.6 Interpretation of relevant genes

To investigate the relevant genes, gene set enrichment analysis (GSEA) [11] is used, implemented with GSEAPY [14]. GSEA tests a ranked list of genes against known gene sets to locate biological meaning. Preranked GSEA, which is the specific variant used in this work, computes a running sum enrichment score (ES) over a ranked list of all genes available, weighted by the magnitude of their rank metric. It then assesses this against a null distribution constructed by gene-set permutation, essentially comparing randomly selected parts of the ranking and comparing them with the test set.

Genes were ranked using their DxC_{462} decoder loadings. This ranking was then inspected manually, followed by testing against two gene set sources. The first source is the Reactome Pathway Knowledgebase [12] which is used to describe the biological program expressed by the factor. Secondly, the set of genes linked to the SNPs of the RA GWAS [9], were used, which was done as this is the set our PRS is based on. To investigate the contribution of the HLA region, the latter source was run both on the full set, and on a set with HLA genes removed. Following GSEA standard practices [11], an FDR threshold of $q < 0.25$ was used.

6 Responsible Research

This work is conducted within the framework set out in the *TU Delft Vision on Integrity 2018-2024*[15]. The following section reflects on the ethical aspects of the research and the reproducibility of methods, focusing on the following pillars: reproducibility, replicability, use of sensitive data, fairness in data and algorithms, data storage and ownership.

6.1 Reproducibility & Replicability

Reproducibility and Replicability, these being the ability of others to obtain the same results by analysing the same data and the ability of others to obtain consistent results by analysing new data with the same methodologies respectively, are crucial for responsible research. In our case these are ensured by first and foremost being transparent about the data used, code availability through a GitHub repository[1], and thorough documentation of all procedures followed, including potential hyperparameters, random seeds and evaluation procedures. If full reproduction is not feasible, this is made clear in documentation and not obscured. The report is additionally made publicly accessible to the TU Delft community through the institutional repository.

6.2 Use of sensitive data

Part of the data used in this work are considered sensitive. This stems from the data being taken from donors and including personally identifying information, medical or otherwise. These are covered by a Controlled-Access Data Use Certificate which restricts redistribution of data and limits what can be shared in this report and the accompanying repository. What this means concretely is that raw data is not included in any publicly accessible medium, such as tables in this analysis, and only processed versions are included, where sensitive data is aggregated or in some other way obfuscated. While this may somewhat limit reproducibility, it does not stop it as use of data can be requested by other researchers [13].

6.3 Fairness in data and algorithms

Part of responsible research is ensuring protected attributes of people are handled well during any implementation and analysis. In this work, features like age, sex and genetic ancestry are included as covariates in analysis in order to control for their effects on the outcome of interest and to prevent bias.

6.4 Data storage & Ownership

The bulk of the data used for this work, including raw and processed data along with the code itself, are stored in the Delft AI Cluster (DAIC). This TU Delft managed infrastructure provides access control, appropriate for use of sensitive data. As described above, the code is additionally stored on a GitHub repository, with data explicitly excluded from version control. Ownership of the data, code and derived results has been discussed and clarified in consultation with the supervisors.

[1] https://github.com/tonitsou/Associating-single-cell-latent-factors-with-genetic-risk

6.5 Use of Large Language Models

A large language model (LLM) was used in this work to assist with certain tasks like formatting latex, restructuring small excerpts, creating menial code and debugging. Every piece of generated information was checked critically before integration into the work. No sensitive data was shared with LLMs at any point in the project. Representative prompts can be found in Appendix E.

References

[1] J. Kurkó, T. Besenyei, J. Laki, T. T. Glant, K. Mikecz, and Z. Szekanecz, "Genetics of rheumatoid arthritis — a comprehensive review," *Clinical Reviews in Allergy & Immunology*, vol. 45, no. 2, pp. 170–179, 2013.

[2] A. Gerussi, B. Soskic, R. Asselta, P. Invernizzi, and M. E. Gershwin, "GWAS and autoimmunity: What have we learned and what next," *Journal of Autoimmunity*, vol. 133, p. 102922, 2022.

[3] C. Liu, R. Joehanes, J. Ma, Y. Wang, X. Sun, A. Keshawarz *et al.*, "Whole genome DNA and RNA sequencing of whole blood elucidates the genetic architecture of gene expression underlying a wide range of diseases," *Scientific Reports*, vol. 12, no. 1, p. 20167, 2022.

[4] D. Vagiaki, T. Heinen, M. Saraswat, B. Clarke, and O. Stegle, "Mapping trans-eQTLs at single-cell resolution using Latent Interaction Variational Inference," bioRxiv preprint, 2026, https://www.biorxiv.org/content/10.1101/2026.02.04.703363.

[5] T. Konuma and Y. Okada, "Statistical genetics and polygenic risk score for precision medicine," *Inflammation and Regeneration*, vol. 41, no. 1, p. 18, 2021.

[6] I. J. Kullo, "Clinical use of polygenic risk scores: current status, barriers and future directions," *Nature Reviews Genetics*, vol. 27, no. 4, pp. 246–263, 2026.

[7] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2013.

[8] ——, "An introduction to variational autoencoders," *Foundations and Trends in Machine Learning*, vol. 12, no. 4, pp. 307–392, 2019.

[9] K. Ishigaki, S. Sakaue, C. Terao, Y. Luo, K. Sonehara, K. Yamaguchi *et al.*, "Multi-ancestry genome-wide association analyses identify novel genetic mechanisms in rheumatoid arthritis," *Nature Genetics*, vol. 54, no. 11, pp. 1640–1651, 2022.

[10] F. Privé, B. J. Vilhjálmsson, H. Aschard, and M. G. B. Blum, "Making the most of clumping and thresholding for polygenic scores," *The American Journal of Human Genetics*, vol. 105, no. 6, pp. 1213–1221, 2019.

[11] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette *et al.*, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences*, vol. 102, no. 43, pp. 15 545–15 550, 2005.

[12] M. Gillespie, B. Jassal, R. Stephan, M. Milacic, K. Rothfels, A. Senff-Ribeiro *et al.*, "The reactome pathway knowledgebase 2022," *Nucleic Acids Research*, vol. 50, no. D1, pp. D687–D692, 2022.

[13] F. Zhang, A. H. Jonsson, A. Nathan, N. Millard, M. Curtis, Q. Xiao *et al.*, "Deconstruction of rheumatoid arthritis synovium defines inflammatory subtypes," *Nature*, vol. 623, no. 7987, pp. 616–624, 2023.

[14] Z. Fang, X. Liu, and G. Peltz, "GSEApy: a comprehensive package for performing gene set enrichment analysis in Python," *Bioinformatics*, vol. 39, no. 1, p. btac757, 2023.

[15] Delft University of Technology, "TU Delft vision on integrity 2018–2024," Committee Reassessment Integrity Policy, TU Delft, 2018, https://www.tudelft.nl/en/about-tu-delft/strategy/integrity-policy/tu-delft-vision-on-integrity-2018-2024.

A List of different PRSs

Abbreviation	Full name
RA	Rheumatoid arthritis
PsA	Psoriatic arthritis
Ax-SpA	Axial spondyloarthritis
JIA	Juvenile idiopathic arthritis
SLE	Systemic lupus erythematosus
Sjogren's	Sjögren's syndrome
Graves	Graves' disease
Hashimoto	Hashimoto's thyroiditis
OA	Osteoarthritis (any site)
OA-finger	Osteoarthritis of the finger
OA-hand	Osteoarthritis of the hand
OA-hip	Osteoarthritis of the hip
OA-hiporknee	Osteoarthritis of the hip or knee
OA-knee	Osteoarthritis of the knee
OA-spine	Osteoarthritis of the spine
OA-thumb	Osteoarthritis of the thumb
Gout	Gout
BMI	Body mass index
Smoking	Smoking
UC	Ulcerative colitis
Crohn's	Crohn's disease

Table 4: A table of all PRS scores calculated and used in our analysis

B OA analysis - D329

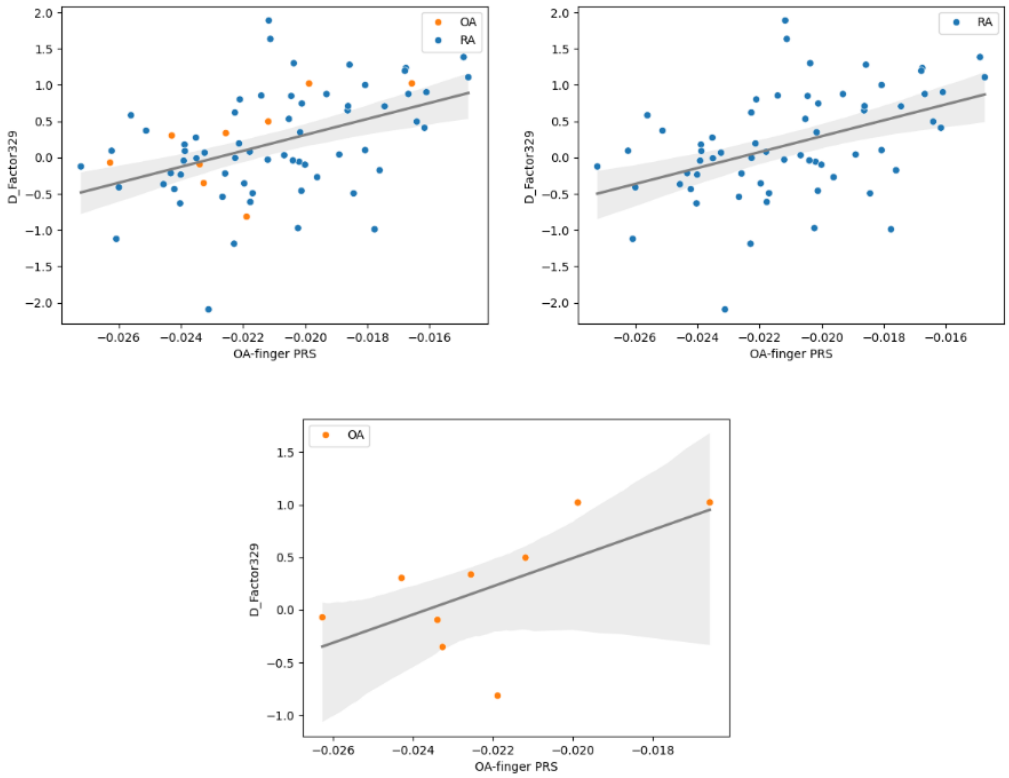

Figure 10: Scatterplots of D_{329}, the nominally significant D factor for OA-finger. The conditions are the following: OA-finger PRS on pooled cohort (Left), OA-finger PRS on RA cohort (Right) and OA-finger PRS on OA cohort. OA donors are indicated in orange.

C Expanded genetic pc plot

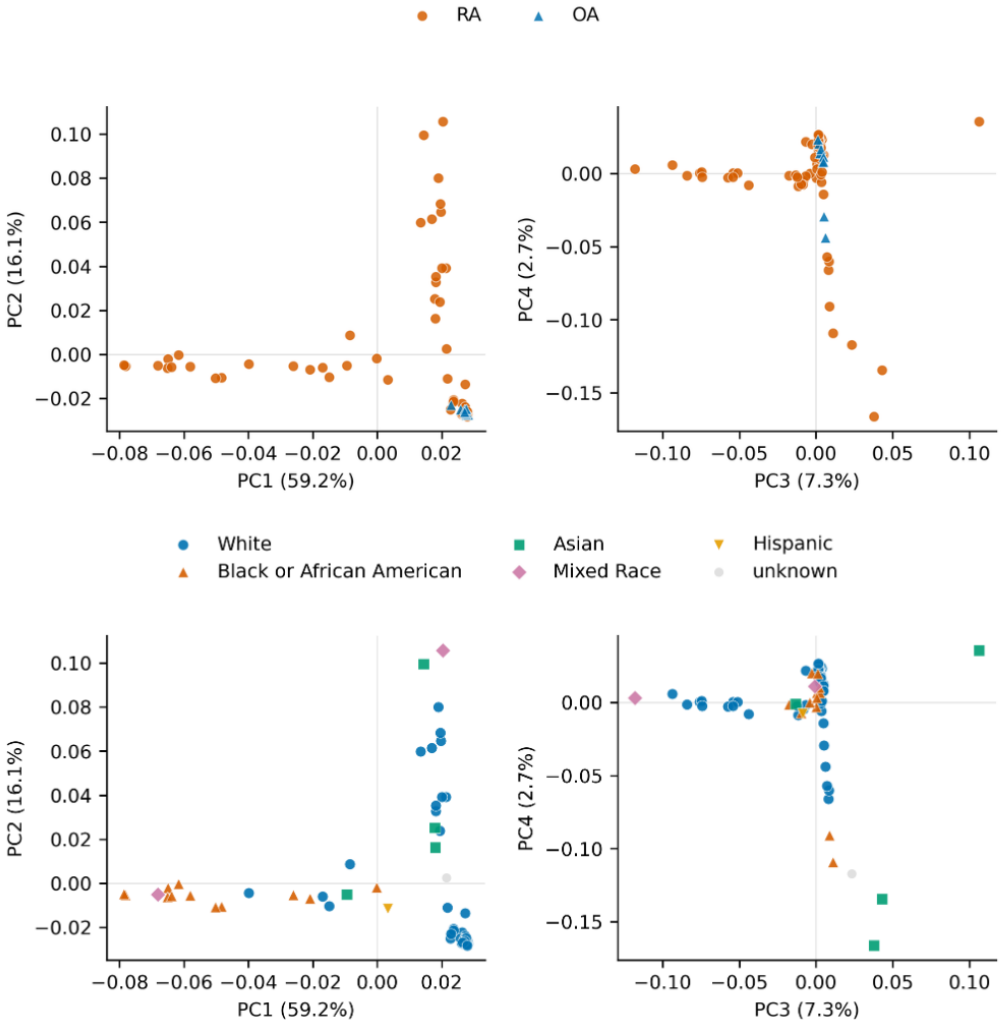

Figure 11: Scatterplots of first four genetic principal components (which explain approximately 85, 3% of the total variance). Additionally donors are annotated according to their disease (top) and race (bottom) which acts as a proxy for ancestry.

D D462 loading strip

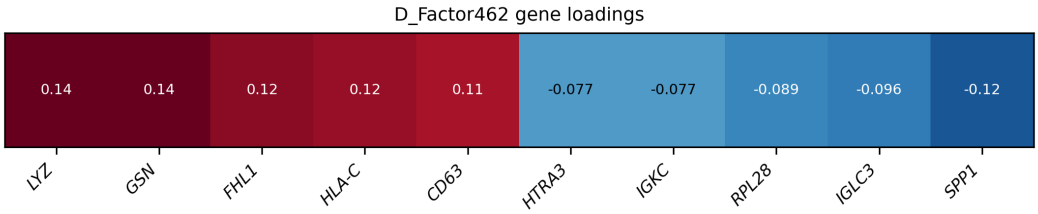

Figure 12: Head and tail genes of D_{462} ordered by loading. The rest of the genes lie between CD63 and HTRA3 in rank.

E LLM indicative prompts

- "Create a properly formatted latex table with the following headers and row names, leave placeholders for data."
- "Write a short script that joins these tables on their shared identifier columns and drops duplicate rows."
- "How can I merge the following plots into a joint one and move their respective legends above them?"
- "Write a small function that parses a delimited column into a clean list, splitting only on the spaced separator so hyphenated names are preserved."
- "Please parse the following error message and explain what could be causing it."
- "How would you weave in this sentence into the following paragraph?"