# Characterizing the temporal evolution of alloreactive T cell clonotypes using probabilistic basis decomposition

**Anonymous Authors**[1]

## Abstract

Extensive work has been performed in modeling temporal dynamics of gene expression to gain a better understanding of underlying regulatory mechanisms. However, there is a scarcity of computational frameworks for studying the evolution of the immune repertoires from time-series data. Here, we adapt a previous probabilistic basis decomposition technique (Nazaret A, 2022) to study T cell clone dynamics in bone marrow transplant recipient patients with graft-vs-host-disease (GVHD). Our analysis reveals a different temporal dynamic between alloreactive and nonalloreactive clones, and presents a framework for quantifying dynamics associated with clinical groups.

## 1. Introduction

The rapid development of high-throughput and single-cell genomic sequencing technologies has enabled recent advances in the temporal analysis of single-cell transcriptomic data, which reveals novel insights into cell differentiation and state transition. These include computational approaches for inferring pseudotime or trajectories of cells (Cannoodt et al., 2016; Lange et al., 2022; Setty et al., 2016; Trapnell et al., 2014). At the same time, technological development also brought forth a revolution in the characterization of immune repertoires. In particular, studying the diversity of the T cell repertoire, mainly generated by somatic rearrangements of the T cell receptor (TCR) $\alpha, \beta$ chains, has important clinical and biological implications, such as in prognostic predictions in cancer (Aran et al., 2022). However, it is now well-known that the composition of immune cell states evolves with time. We therefore need new computational frameworks for tracking the temporal dynamics of T-cell clonotypes defined as unique antigen-specific nucleotide sequences arising from rearrangement. In particular, T cell proliferation is activated by clonal selection resulting in a sufficient number of antigen-specific T cells with the same clonotype, mounting an effective immune response.

Conventional analysis of TCR repertoires employs statistical metrics from ecology. Current analysis tools for single-cell T cell clonotypes assess clonotype frequencies and diversities. These methods include cdr3tools(Zhang et al., 2020), immunarch(ImmunoMind Team, 2019), and scirpy(Sturm et al., 2020) for single-cell TCR analysis. Numerous methods have been developed to characterize the differences between samples or perform clustering on TCR. TCRNET (Ritvo et al., 2018) uses sequences similarity to cluster similar TCRs, and TCRdist (Dash et al., 2017) evaluates differences between TCRs, whereas RECOLD (Yokota et al., 2017) studies the differences between samples. On the other hand, Davidsen et al. (Davidsen et al., 2019) proposes a Variational Encoder based generative model by treating TCR sequence formation as string generation, and DeepTCR(Sidhom et al., 2021) solves classification problems using VAE generative modeling. However, there are very few attempts to characterize the temporal dynamics of TCR data.

In this work, we adapt and deploy a scalable basis decomposition approach (Nazaret A, 2022) to characterize temporal clonotype patterns similar to gene expression. One advantage of this approach is that it does not make assumptions about the form of the patterns across all clones. Like genes, clonotypes can be represented as a combination of one or more representative patterns. Most importantly, basis decomposition allows interpretable comparisons between conditions or clinical groups with consideration of temporal dynamics.

We establish an application of this computational framework in studying transplantation, where T cell clonality is especially important because T cells are the major cause of graft rejection, particularly alloreactive T cells (DeWolf et al., 2018). These are T cells that encounter antigens unseen in donor thymic development, thus those T cells from the donor will attack the recipient's tissue, which leads to GVHD. Symptoms of GVHD mostly show up in the

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by *the 2023 ICML Workshop on Computational Biology*. Do not distribute.

skin and gut tissues. Experimentally, alloreactive clones can be discovered with TCR sequencing and a standard *in vitro* functional assay, the mixed lymphocyte reaction (MLR). MLR involves the co-culture of both donor and recipient's PBMCs (Peripheral Blood Mononuclear Cells) and sequencing of expanded donor T cells. Our hypothesis is that alloreactive T cell clones are defined by representative patterns, i.e. basis functions, that are different from non-alloreactive T cell clones. Also, repertoires of patients with different grades of GVHD could exhibit varying patterns. Specifically, we aim to identify alloreactive fingerprints and patterns of alloreactive clones in the blood with TCRB Sequencing and their correlation with clinical and pathological GVHD with basis decomposition.

## 2. Methods

### 2.1. Data

Patients at Columbia University Medical Center with acute myeloid leukemia received bone marrow transplantation from haplo and matched unrelated donors. Mixed lymphocyte reactions (MLRs) were carried out with PBMCs from both donor and recipient. Expanded T cells were sent for TCRB Adaptive Immunosequencing, and patients' blood samples were repeatedly collected at each doctor's visit. 12 patients samples were profiled with a median duration of 272 days and 8 time points per patient.

### 2.2. Preprocessing

Alloreactive clones (unique CDR3 sequences on the TCR chain) were defined as clones with frequencies twice as much as their frequencies in donor and higher than 0.0001 in the MLR expanded population. Out of alloreactive clones, we also defined expanded alloreactive clones by finding the elbow point of the smoothed clonotype frequency curve (using local polynomial regression). Any clonotypes that existed in fewer than two-time points were filtered out. Patients with sample collections that stopped before 180 days were removed. Due to limited number of timepoints and T cells detected with immunosequencing, the time-series data per clonotype is extremely sparse. To tackle this sparsity, we performed a Gaussian Process regression to achieve a smooth model of temporal dynamics per clonotype, while accounting for dependencies between time points. The RBF kernel was used for proper fitting of the clonotype data.

### 2.3. Basis decomposition

For the basis decomposition model, the input data of the model consists of frequencies of $G$ clonotypes observed over time in $C$ conditions or patient groups, e.g. GVHD or healthy. For clonotype $g$ in condition $c$, the model decomposes data point $u_g, c$ with a latent vector $\beta_g, c$ of K dimensions. Each dimension $k$ represents a latent basis pattern with coefficient $\beta_g, c, k$. The weights $\beta_g, c, k$, and latent bases generate clonotype patterns, $\mathrm{u}_g, c$, and the weights are positive values drawn from a method mimicking non-negative matrix factorization (Nazaret A, 2022). Weights $\beta_g, c, k$ are independently drawn from an exponential distribution with mean $\eta, (\varepsilon(\eta))$, with density

$$\rho(\beta|\eta) = \eta * e^{-\eta\beta} 1(\beta \geq 0)$$

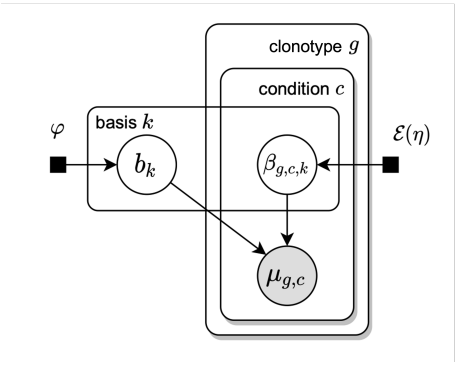


*Figure 1.* Graphical model for the basis decomposition model (Nazaret A, 2022) with clonotype frequency as input.

The basis functions $b$ were sampled by drawing neural network parameters ($\theta$, 1D neural network with two hidden layers of 32 units each with tanh activation). Variational inference was used to approximate the exact posterior, by defining a family of approximate distributions $Q$ over latent variables and finding the member of this family that is closest to the exact posterior. Then, the posterior inference becomes an optimization problem. Mean-field approximations for the variational family were utilized, and the variational distribution is factorized as:

$$q(\theta, \beta) = \prod_{k,j} q(\theta_k, j) \prod_{g,c,k} q(\beta_g, c, k)$$

ELBO (evidence lower bound) is used to approximate the posterior.

Since the model does not handle negative values, the frequency values were shifted by the minimum values from Gaussian Processed data, and normalized to the maximum value in the dataset. The number of basis functions used was 5 (to ensure proper fitting and improve identifiability), and the l1 regularization of the coefficient $\beta$ has a strength of 10, leading to sparsity and improved interpretability.

## 3. Results

We investigated the proportion of unique alloreactive clones out of all unique clones, clone diversity, and accumulative frequencies of alloreactive clones over time. Our goal is to identify alloreactive fingerprints, patterns of alloreactivity over time in the blood with TCRB sequencing, and its correlation with the grade of GVHD and drug treatments. Thus,

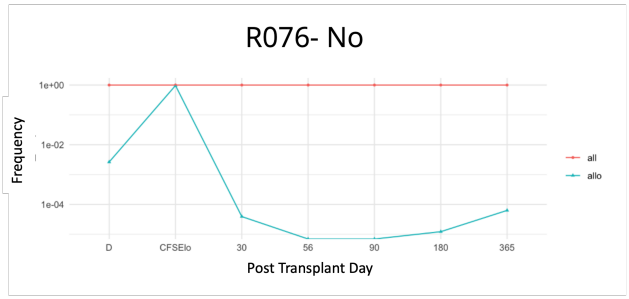


*Figure 2.* Accumulative frequency of alloreactive clones for a patient who did not develop GVHD.

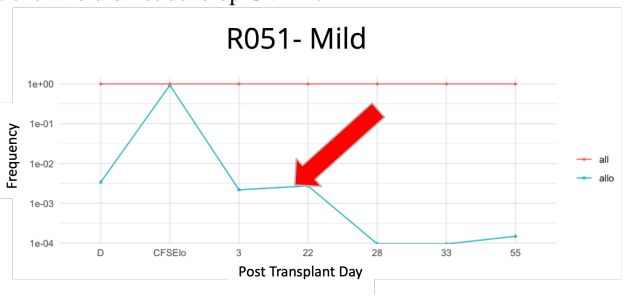


*Figure 3.* Accumulative frequency of alloreactive clones for a mild GVHD patient. The red arrow indicates the onset time point of GVHD.

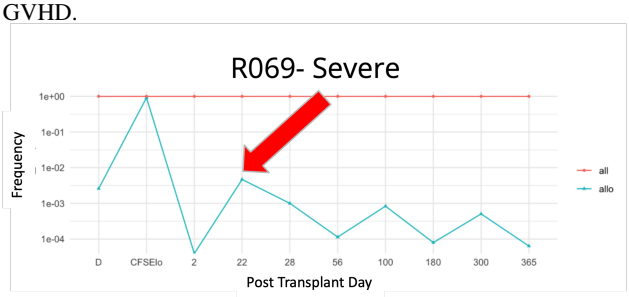


*Figure 4.* Accumulative frequency of alloreactive clones for a severe GVHD patient.The red arrow indicates the onset time point of GVHD from clinical diagnosis.

blood samples and MLR were collected and performed from 12 patients with different grades of GVHD and drug treatments.

### 3.1. Accumulative frequency of alloreactive clones

We first performed an unbiased analysis of temporal patterns in the accumulative frequency of alloreactive clones associated with GVHD. Strikingly, we found that the onset of GVHD occurs when the slope of accumulative frequency increases. Interestingly, the slope of increase in alloreactive clone frequency is steeper in patients with more severe GVHD (Fig. 2-4). More severe GVHD also coincided with periodic increases in accumulative frequency. This data thus suggests greater expansion of the alloreactive clones compared to non-alloreactive clones, leading to the onset of GVHD. These findings thus motivate modeling the temporal dynamics of expanded and individual antigen-specific clonotypes in the next two subsections.

### 3.2. Expanded alloreactive clones

To explore the temporal patterns of inidividual alloreactive T cell clones, we first studied the expanded alloreactive clones. As mentioned in the Methods section, the expanded alloreactive T cell clones were defined according to a sample-specific threshold. We first studied expanded alloreactive T cell clones over time by grouping patients into No GVHD, Mild GVHD, and Severe GVHD (Appendix Fig.12-14), and we found that more severe patients have more expanded alloreactive clones over time, whereas mild patients has less alloreactive clones. However, there is an outlier on a No GVHD patients.

Furthermore, we investigated the effect of drug treatment (PTCy) on expanded alloreactive T cell clones, which is expected to reduce the rate of GVHD for haplo transplantation(Ritacco et al., 2023). Here, we discovered that PTCy treatment indeed reduces the rate of expanded alloreactive T cell clones (Appendix Fig. 15-16) (in 8 out of 9 patients).These collective analyses on T cells inspired us to look further into temporal dynamics of individual T cell clonotypes.

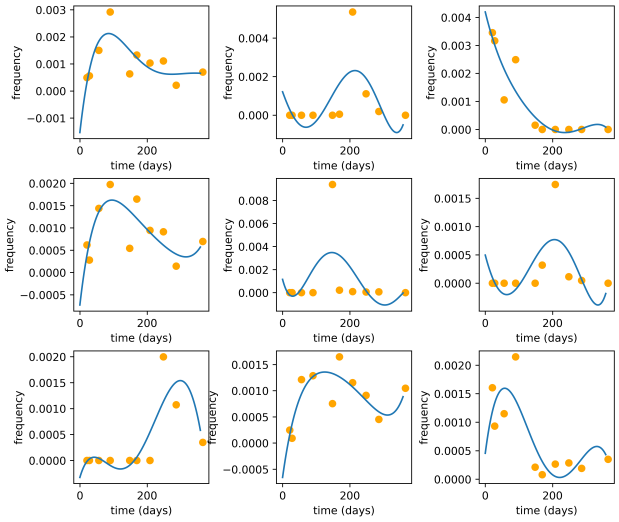


*Figure 5.* Gaussian Process (GP) regression fit to representative clonotypes (orange dots); GP mean is shown in blue.

### 3.3. Basis decomposition of T cell clonotype dynamics

We then investigated how temporal patterns of alloreactive clones might differ across different conditions. We first preprocessed the sparse clone-level data with the Gaussian process model. Then, using the basis decomposition model (Fig. 5), we inferred 5 basis functions (Appendix Fig.15) that are capable of reconstructing the clonotype fre-

quency data, which indicates proper fitting of the data to the model(Appendix Fig. 16). Recapitulating clonotype information with basis functions presents an exciting framework for interpreting dominant temporal patterns and their association with clinical information. Interestingly though, we did not find a significant difference in decomposition weights between patients with and without GVHD (Appendix Fig. 17).

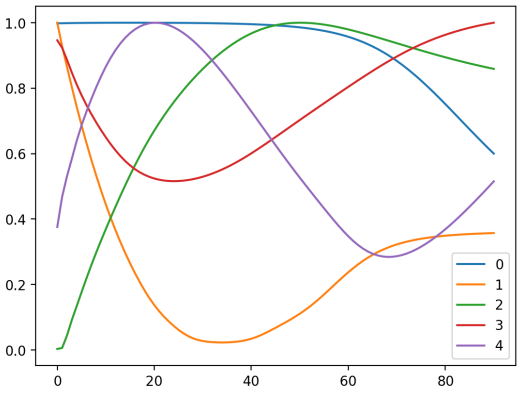


*Figure 6.* Basis functions for alloreactive and non-alloreactive clones .

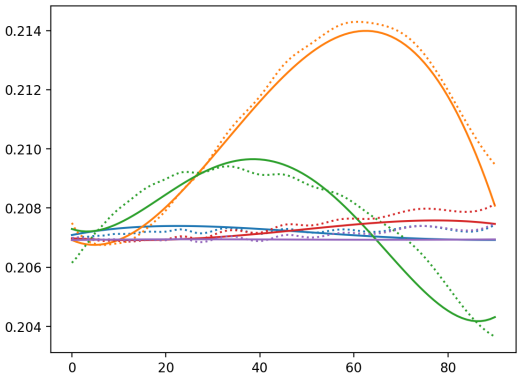


*Figure 7.* Reconstructed clonotype data (dotted) and real clonotype data (solid).

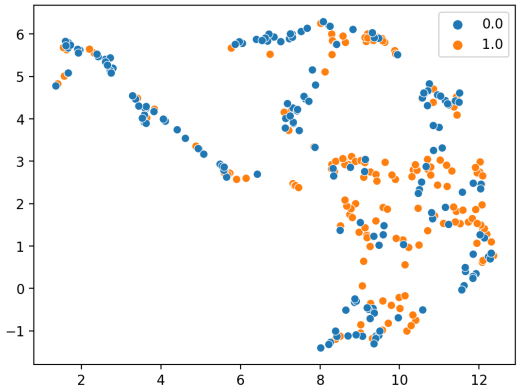


*Figure 8.* Umap projection of weights of basis function of clonotypes. 0 = non-alloreactive, 1 = alloreactive.

We thus investigated the difference between alloreactive clones and non-alloreactive clones using our pipeline. We accounted for sample size bias by subsampling both non-alloreactive and alloreactive clones to the a similar number (162 alloreactive clones and 170 non-alloreactive clones respectively). Fig. 6 shows the inferred basis functions. In addition, we show accurate reconstruction of data with the basis decomposition model (Fig. 7). We then visualized the weights of the basis functions on a 2D UMAP, where we found a decent separation between non-alloreactive clones versus alloreactive clones (with upper two branches mainly composed of non-allreactives) (Fig. 8). Histograms were plotted to visualize the difference in weights of basis function between alloreactive versus non-alloreactive clones. Basis function 0 exhibits the most distinct difference, with p-values of 0.005 from t-test. Basis 0 in Fig.6 shows persistent pattern until day 60+, indicating longer-term persistence of alloreactive clones. It is not surprising that some dynamics are shared between the two groups as there are indeed alloreactive clones disappearing over time and non-alloreactive clones could be activated/expanded over time due to encountering of other types antigens.

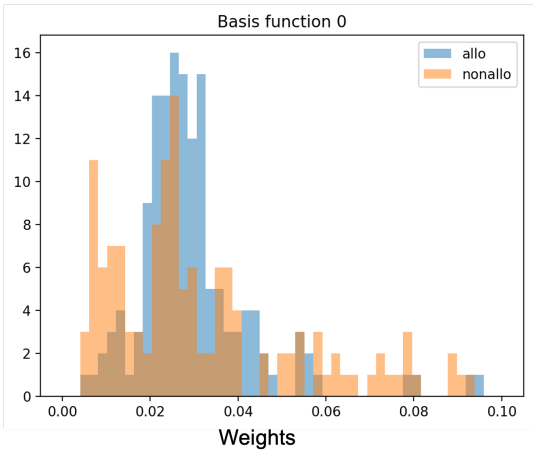


*Figure 9.* Histograms for representative basis function with differential weights between alloreactive and non-alloreactive clones (tails truncated for visualization).

## 4. Conclusions and Future Directions

In summary, our work has demonstrated the feasibility of applying a probabilistic basis decomposition model to study T cell clonotype dynamics with interpretable comparison. We are the first to our knowledge to model T cell clonotype datasets to track temporal dynamics via representative patterns. We are able to identify representative patterns dominant in alloreactive vs. non-alloreactive clones. This preliminary work will lay a foundation to continue investigating TCR temporal dynamics and integrate the Gaussian Process in the basis decomposition model for joint inference. Furthermore, we aim to integrate the frequency of clonotypes with the unique cdr3 sequences and VDJ recom-

bination as well as antigen-specificity datasets, and perform simultaneous clustering of T cell clonotypes in future work.

## References


Aran, A., Garrigós, L., Curigliano, G., Cortés, J., and Martí, M. Evaluation of the TCR repertoire as a predictive and prognostic biomarker in cancer: Diversity or clonality? *Cancers*, 14(7), March 2022.

Cannoodt, R., Saelens, W., Sichien, D., Tavernier, S., Janssens, S., Guilliams, M., Lambrecht, B., Preter, K. D., and Saeys, Y. Scorpius improves trajectory inference and identifies novel modules in dendritic cell development. *bioRxiv*, 2016. doi: 10.1101/079509. URL https://www.biorxiv.org/content/early/2016/10/07/079509.

Dash, P., Fiore-Gartland, A. J., Hertz, T., Wang, G. C., Sharma, S., Souquette, A., Crawford, J. C., Clemens, E. B., Nguyen, T. H. O., Kedzierska, K., La Gruta, N. L., Bradley, P., and Thomas, P. G. Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature*, 547(7661), July 2017.

Davidsen, K., Olson, B. J., DeWitt, III, W. S., Feng, J., Harkins, E., Bradley, P., and Matsen, IV, F. A. Deep generative models for T cell receptor protein sequences. *Elife*, 8, 2019.

DeWolf, S., Grinshpun, B., Savage, T., Lau, S. P., Obradovic, A., Shonts, B., Yang, S., Morris, H., Zuber, J., Winchester, R., Sykes, M., and Shen, Y. Quantifying size and diversity of the human T cell alloresponse. *JCI Insight*, 3(15), August 2018.

ImmunoMind Team. immunarch: An R Package for Painless Bioinformatics Analysis of T-Cell and B-Cell Immune Repertoires, August 2019. URL https://doi.org/10.5281/zenodo.3367200.

Lange, M., Bergen, V., Klein, M., Setty, M., Reuter, B., Bakhti, M., Lickert, H., Ansari, M., Schniering, J., Schiller, H. B., Pe'er, D., and Theis, F. J. CellRank for directed single-cell fate mapping. *Nat. Methods*, 19 (2):159–170, February 2022.

Nazaret A, Fan JL, P. D. A. E. Probabilistic basis decomposition for characterizing temporal dynamics of gene expression. 2022.

Ritacco, C., Köse, M. C., Courtois, J., Canti, L., Beguin, C., Dubois, S., Vandenhove, B., Servais, S., Caers, J., Beguin, Y., Ehx, G., and Baron, F. Post-transplant cyclophosphamide prevents xenogeneic graft-versus-host disease while depleting proliferating regulatory T cells. *iScience*, 26(3):106085, March 2023.

Ritvo, P.-G., Saadawi, A., Barennes, P., Quiniou, V., Chaara, W., El Soufi, K., Bonnet, B., Six, A., Shugay, M., Mariotti-Ferrandiz, E., and Klatzmann, D. High-resolution repertoire analysis reveals a major bystander activation of tfh and tfr cells. *Proc. Natl. Acad. Sci. U. S. A.*, 115(38):9604, September 2018.

Setty, M., Tadmor, M. D., Reich-Zeliger, S., Angel, O., Salame, T. M., Kathail, P., Choi, K., Bendall, S., Friedman, N., and Pe'er, D. Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nature Biotechnology*, may 2016. doi: doi:10.1038/nbt.3569. URL http://dx.doi.org/10.1038/nbt.3569.

Sidhom, J.-W., Larman, H. B., Pardoll, D. M., and Baras, A. S. DeepTCR is a deep learning framework for revealing sequence concepts within t-cell repertoires. *Nat. Commun.*, 12(1):1605, March 2021.

Sturm, G., Szabo, T., Fotakis, G., Haider, M., Rieder, D., Trajanoski, Z., and Finotello, F. Scirpy: a scanpy extension for analyzing single-cell t-cell receptor-sequencing data. *Bioinformatics*, 36(18):4817–4818, September 2020.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S., and Rinn, J. L. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.*, 32(4):381–386, April 2014.

Yokota, R., Kaminaga, Y., and Kobayashi, T. J. Quantification of Inter-Sample differences in T-Cell receptor repertoires using Sequence-Based information. *Front. Immunol.*, 8, 2017.

Zhang, Y., Yang, X., Zhang, Y., Zhang, Y., Wang, M., Ou, J. X., Zhu, Y., Zeng, H., Wu, J., Lan, C., Zhou, H.-W., Yang, W., and Zhang, Z. Tools for fundamental analysis functions of TCR repertoires: a systematic comparison. *Brief. Bioinform.*, 21(5):1706–1716, September 2020.

# A. Appendix

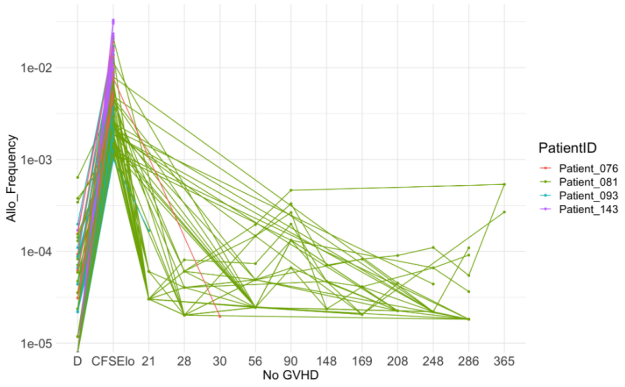


*Figure 10.* Expanded alloreactive clones for no GVHD patients (with transplantation).

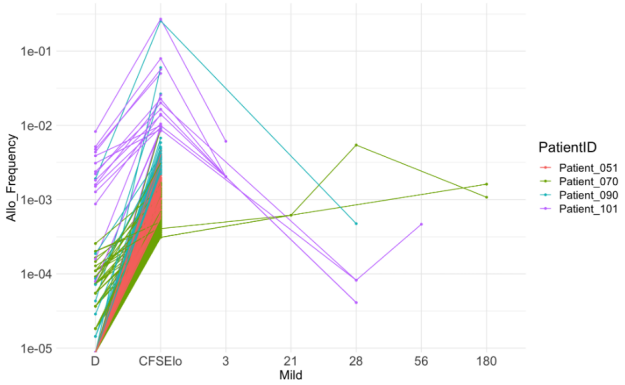


*Figure 11.* Expanded alloreactive clones for mild GVHD patients.

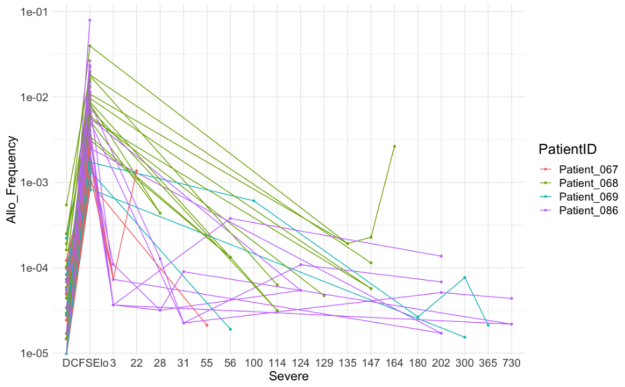


*Figure 12.* Expanded alloreactive clones for severe GVHD patients.

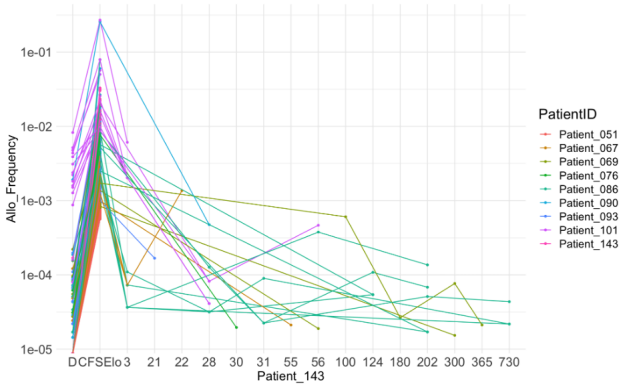


*Figure 13.* Expanded alloreactive clones for PTCy treated patients.

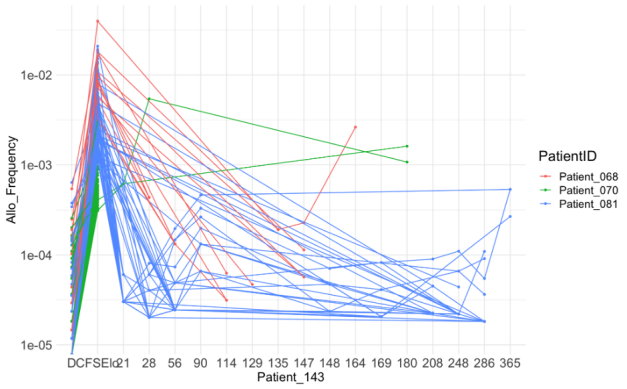


*Figure 14.* Expanded alloreactive clones for PTCy untreated patients.

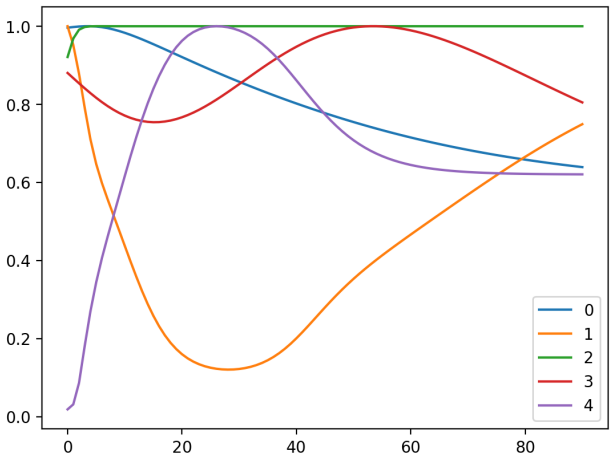


*Figure 15.* Inferred basis functions for alloreactive clones (all patients).

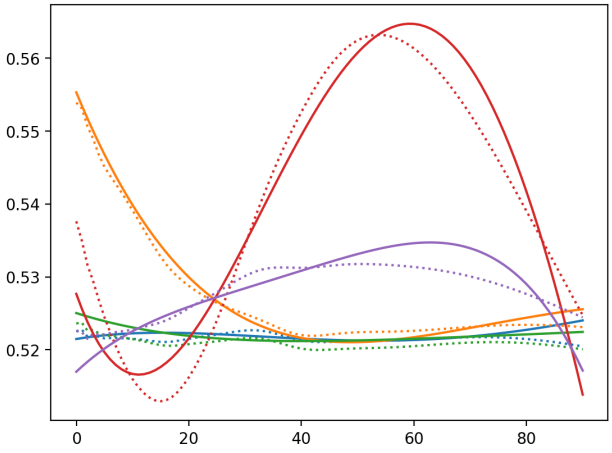


*Figure 16.* Reconstructed clonotype data (dotted) from basis decomposition and clonotype data (solid) smoothed with a GP for example clonotypes.

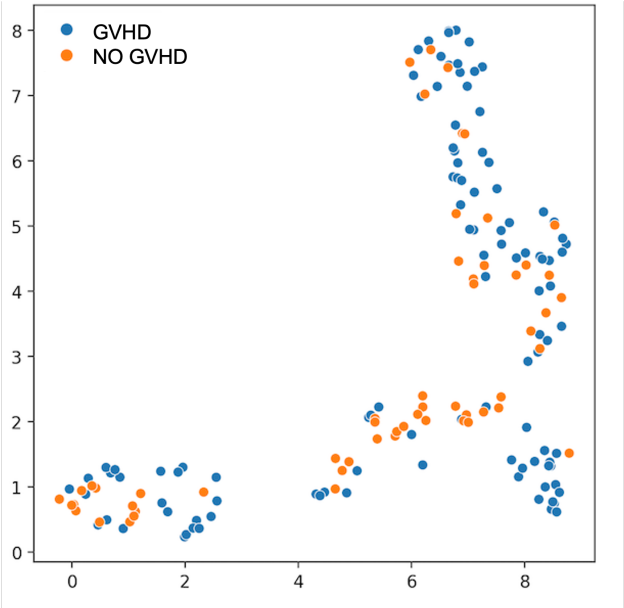


*Figure 17.* Umap projection of weights of basis function of alloreactive clones.