Variance component mixture modelling for longitudinal T-cell receptor clonal dynamics

David Swanson[1], Alexander Sherry[2], and Chad Tang[2]

[1]Department of Biostatistics,
[2]Department of Radiation Oncology,
University of Texas MD Anderson Cancer Center

January 21, 2025

Abstract

Studies of T cells and their clonally unique receptors, which mediate T cell function through highly specific binding of target epitopes, have shown promise in elucidating the association between immune response and human disease, from cancer and infectious disease to autoimmune conditions. Methods to identify T-cell receptor clones which expand or contract in response to certain conditions or therapeutic strategies have so far been limited to longitudinal pairwise comparisons of changes in clone frequency with multiplicity adjustment. Mixture models have potential to play a useful role in distinguishing dynamic T-cell receptor clones from static ones, allowing one to relate these features to exposures and different markers of health and prognosis. While it is common to mix on the location or scale parameter of a family of distributions, mixing on the parameterization of the model is exploited in our approach, with one mixture component allowing TCR clones to have variable mean over longitudinal followup, whereas the other component does not. Our model leverages Bayesian hierarchy, utilizing the posterior predictive distribution of Gamma-Poisson conjugacy to arrive at negative binomial and negative multinomial distributions of components, each modified according to an offset term for receptor read count normalization. An EM-algorithm is developed to estimate hyperparameters, and validity of the approach is demonstrated in simulation. The model identifies a statistically significant and clinically relevant increase in TCR clonal dynamism among metastasis-directed radiation therapy in a cohort of prostate cancer patients.

1 Introduction

Immune response plays a central role in human health and addressing the many foreign agents people are regularly exposed to. T-cells are an important piece of one's adaptive immune system, in part by recognizing antigens with receptors that exhibit an exceptional degree of diversity via highly variable gene recombination in portions of the receptor [15, 8]. Study of T-cell receptor (TCR) repertoires has gained attention in recent years because immunosequencing receptors allows one to examine TCR clonal dynamics and their association with biomarkers, disease, and other aspects of human biology [25, 21, 17]. Clones of receptors binding an antigen tend to expand, which can be measured by calculating the proportion of a certain clone within the repertoire using sequencing technologies and examining the quantity over time [7]. Other clones contract or are relatively static proportionally within the repertoire, and these dynamics have been looked at longitudinally [11]. Contraction of clones, in particular, often occurs once the antigenic stimulus is withdrawn, or if TCR binding is ineffective against the dominant epitopes present in the microenvironment. Identifying expanding and contracting clones has taken different forms in the past, though a popular approach places the problem within a pairwise time comparison and multiple testing framework, identifying significant changes in clonal proportions over two points in time within a Beta-Binomial conjugate prior setting which effectively moderates estimated rate changes by pulling clones with low read counts toward the null hypothesis of no change [22]. The procedure then applies false discovery proportion control to address the severe multiple testing burden involved in the analysis [3]. The approach has been successful at identifying expanding and contracting clones and relating the dynamics to a variety of diseases. Cancer has been a disease of primary interest with respect to longitudinal TCR clonal movement and trajectories and their associations with treatment and prognosis, with recent studies demonstrating clonal expansions and contractions significantly associated with radiation therapy directed at metastases [24].

However, routine implementation of TCR sequencing in clinical practice and management of solid tumors has been limited to date, and particular challenges include heterogeneity and uncertainties in modeling approaches. Comprehensive statistical modeling approaches to understanding longitudinal T-cell receptor clonal dynamics have so far not been pursued in the literature to our knowledge, which might account for variable follow-up by observation, missingness, and error term specification so that what constitutes significant clonal frequency change is explicit. Several challenges arise when trying to develop a model for these dynamics. Because of heterogeneity in the total number of template reads within the biological unit being interrogated, one models a fraction, rather than an absolute count, since it is the clonal fraction relative to others that elucidates biological fluctuation and therefore interest. Second, because of extreme antigenic and therefore receptor diversity, most clones are not "public", or shared, across the population being modelled, unlike expression of a certain gene or many gene mutations which though varying across observations have a common interpretation in one's sample [2]. Therefore one generally cannot relate any one clone and its expansion or contraction behavior with patient phenotype, rather, one must consider dynamics of the repertoire as a whole. The challenge of relating clonal dynamics with phenotype is particularly profound for systemic circulating T cells vs T cells that have infiltrated into tumor tissue, although the dynamics and functionality of the systemic TCR repertoire is often of considerable and clinical interest, not to mention more easily studied through peripheral blood draws as compared with serial tissue biopsies. Lastly, while it is often feasible to identify change in some proportion from a statistical perspective, identifying biologically relevant changes, which presumably are of greater interest for prognostic models and other uses, is a different and more nuanced question. Inclusion of an error term that is monotonic increasing in the clone's mean parameter may capture biological mechanism best.

We try to address these challenges with an interesting and novel variance component mixture model, which leverages Bayesian conjugacy under two different parameterizations of a variance component model, one component corresponding to dynamic (expanding or contracting) clones, and one component corresponding to static clones [18, 23]. The model posits clonal mean frequencies being sampled from a Gamma prior which we use hierarchically to sample from a Poisson on that clone's mean parameter λ, with offset term corresponding to the total number of reads within the biological unit being modelled. Under the dynamic component, a different clonal mean parameter λ is hypothesized for each longitudinal follow-up, giving flexibility to the within-clone frequency variability over time. In contrast, the same λ parameterizes all follow-up times under the static component, constraining the entirety of that clone's frequencies for the times observed. For both parameterizations, we marginalize out λ, leveraging conjugacy to arrive at a marginal likelihood that is a product of negative binomial probability mass functions (pmf) for the dynamic component clones and the negative multinomial pmf for the static component clones, both of which are modified due to the offset term. We fit α and β with empirical Bayes, and use the expectation-maximization (EM) algorithm to calculate the probability of component membership for each clone, be it dynamic or static, iterating until convergence [10].

In section 2, we derive the model and means of its fitting, grounding it in hierarchical Bayesian modelling and the broad literature on mixture models. In section 3, we examine sensitivity of the model to different sets of parameters, proportion mixing, and powerings and compare with the multiple testing framework in which such analyses are often done. In section 4 we analyze a cohort of prostate cancer patients who are randomized to metastasis-directed therapy or not and the influence of that exposure on TCR clonal dynamics. We conclude in Section 5 with consideration of model extensions and improvement.

2 Methods

Finite mixture models have played a central role in the development of statistical modelling with broad use across data-generating mechanisms where there is belief that subsets of observations arise under a discrete set of different parameters [13, 6, 9]. Since parameters are generally unknown for each subset, or component, nor is component membership known among observations, these models can be difficult to fit and considerable work has been devoted to understanding how and behavior of subsequent estimates. Simulation-based methods like Markov chain monte carlo (MCMC) and other optimization routines have both proved valuable tools in model fitting [19, 20]. It may be most common to see these mixture models mixing on location and scale parameters in exponential families of distributions for clustering data, though more sophisticated mixtures of regressions or mixtures of experts have also been proposed and applied to different modelling settings [5, 14].

It is less common to mix on the parameterization of a model, which is our proposition here. Our mixture consists of two different Gamma-Poisson hierarchies which upon marginalization yields a negative multinomial distribution governed by two hyperparameters α and β, whereas the other component is a product of negative binomial distributions governed by the same parameters, where the number of terms is the follow-up times within clonotype. Both of these variance components are a function of the number of clonotype TCR reads and the total number of reads across clones per person-time. While variance components are typically used in mixed models to account for within-observation correlation and augment fixed effects, here they are useful to partition dynamic clones from static ones, that is those that exhibit significant heterogeneity in clonal proportion across time versus those that do not [23, 26, 1].

We fit the model by first randomly assigning $E[D_{ij}]$, optimizing with respect to α and β, then iterating until convergence, defined as the sum of changes in component probabilities being less than some ϵ.

2.1 Model framework

Consider the i^{th} clone template read count known from sequencing for person j at time k, C_{ijk}, and the total number of template counts for person j at time k, $O_{jk} = \sum_{i=1}^{U_{jk}} C_{ijk}$, where U_{jk} is the number of unique clones for person j at time k. Since there is significant variability in the total number of template reads and it is of primary biological interest how proportions of clones change over time, we model the proportion C_{ijk}/O_{jk} over time and patient.

We seek to distinguish the "dynamic" clones (denoted $D_{ij} = 1$ for clone i person j), who exhibit significant contraction and/or expansion over longitudinal follow-up, from the static ones ($D_{ij} = 0$). The hierarchical Bayesian formulation envisions that among the static component, clonal mean parameter $\lambda_{ij} \sim \text{Gamma}(\alpha, \beta)$ is invariant over time k and sampled from a Gamma distribution on hyperparameters α and β. Template count reads are assumed Poisson-distributed $C_{ijk} \sim \text{Pois}(\lambda_{ij} O_{jk})$, where the λ_{ij} is scaled by the total number of reads at a person-time O_{jk}, effectively treating its logged value as an offset. In contrast, under the dynamic component, separate λ_{ijk} are assumed sampled for each time point, in which case subsequently $C_{ijk} \sim \text{Pois}(\lambda_{ijk} O_{jk})$. Integration of λ_{ij} and λ_{ijk} yields the respective posterior predictive distributions, which due to conjugacy are here the negative multinomial and negative binomial distributions accordingly. These distributions are appropriate for modelling clonal template counts because of the right-skewed and significant dynamic ranges of this outcome. These distributions are modified according to the O_{jk} via their use as an offset, implicitly modelling clonal proportion dynamics.

2.2 Static model

We have the static component clonal model as

$$C_{ijk}\,|\,(\lambda_{ij}, O_{jk}, D_{ij}=0) \sim \text{Pois}(\lambda_{ij}O_{jk})$$

with

$$\lambda_{ij}|(D_{ij}=0) \sim \text{Gamma}(\alpha, \beta)$$

The likelihood for $C_{ij\cdot} = (C_{ij1}\ldots C_{ijT_j})$ given λ_{ij} and $O_{j\cdot} = (C_{j1}\ldots C_{jT_j})$, $l(C_{ij\cdot}|\lambda_{ij}, O_{j\cdot}) = \prod_k l(C_{ijk}|\lambda_{ij}, O_{jk})$, with T_j the number of time points observed for clone i on person j due to an unbalanced design (ie, missingness of follow up for a subset of patients), is

$$l(C_{ij\cdot}|\lambda_{ij}, O_{j\cdot}) = \prod_{k=1}^{T_j} \frac{\exp(-\lambda_{ij}O_{jk})\,(\lambda_{ij}O_{jk})^{C_{ijk}}}{C_{ijk}!} = \frac{\exp(-\lambda_{ij}\sum_k O_{jk})\lambda_{ij}^{\sum_k C_{ijk}}\prod_{k=1}^{T_j} O_{jk}^{C_{ijk}}}{\prod_{k=1}^{T_j} C_{ijk}!}$$

Since our prior on λ_{ij}, $p(\lambda_{ij}|D_{ij}=0,\alpha,\beta)$, is Gamma$(\alpha,\beta)$, the conditional joint distribution of $(C_{ij\cdot}, \lambda_{ij})\,|\,O_{j\cdot}, D_{ij}=0$ is

$$\begin{aligned} p(C_{ij\cdot}, \lambda_{ij}\,|\,O_{j\cdot}, D_{ij}=0,\alpha,\beta) &= l(C_{ij\cdot}|\lambda_{ij},\, O_{j\cdot})p(\lambda_{ij}|D_{ij}=0,\alpha,\beta) \\ &= \frac{\exp(-\lambda_{ij}\sum_k O_{jk})\lambda_{ij}^{\sum_k C_{ijk}}\prod_{k=1}^{T_j} O_{jk}^{C_{ijk}}}{\prod_{k=1}^{T_j} C_{ijk}!} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda_{ij}^{\alpha-1}\exp\left(-\beta\lambda_{ij}\right) \end{aligned} \tag{1}$$

Collecting terms of equation (1) and leveraging Gamma-Poisson conjugacy, we recognize the kernel of the Gamma on updated parameters $\alpha' = \sum_k C_{ijk} + \alpha$ and $\beta' = \sum_k O_{jk} + \beta$. Then, integrating out λ_{ij}, we obtain the inverse normalizing constant of the Gamma(α', β')

$$\begin{aligned} p(C_{ij\cdot}|O_{j\cdot}, D_{ij}=0,\alpha,\beta) &= \int p(C_{ij\cdot}, \lambda_{ij}|O_{j\cdot}, D_{ij}=0,\alpha,\beta)\,d\lambda_{ij} \\ &= \frac{\beta^\alpha \prod_{k=1}^{T_j} O_{jk}^{C_{ijk}}}{\Gamma(\alpha)\prod_{k=1}^{T_j} C_{ijk}!}\int_0^\infty \lambda_{ij}^{\sum_k C_{ijk}+\alpha-1}\exp\left(-\lambda_{ij}(\beta + \sum_k O_{jk})\right)d\lambda_{ij} \\ &= \frac{\beta^\alpha \prod_{k=1}^{T_j} O_{jk}^{C_{ijk}}}{\Gamma(\alpha)\prod_{k=1}^{T_j} C_{ijk}!} \cdot \frac{\Gamma(\sum_k C_{ijk}+\alpha)}{(\beta + \sum_k O_{jk})^{\sum_k C_{ijk}+\alpha}} \\ &= \frac{\Gamma(\sum_k C_{ijk}+\alpha)}{\Gamma(\alpha)\prod_{k=1}^{T_j} C_{ijk}!}\left(\frac{\beta}{\beta+\sum_k O_{jk}}\right)^\alpha \prod_{k=1}^{T_j}\left(\frac{O_{jk}}{\beta+\sum_k O_{jk}}\right)^{C_{ijk}} \end{aligned} \tag{2}$$

which for discrete α we recognize Equation (2) as a negative multinomial probability mass function on size parameter α, success probability $\beta/(\sum_k O_{jk}+\beta)$, and “offset-normalized” near equal-valued failure probabilities $O_{jk}/(\beta+\sum_k O_{jk})$, $k = 1,\ldots,T_j$, and evaluated at vector $(C_{ij1},\ldots,C_{ijT_j},\alpha-1)$, for clone read counts $C_{ij\cdot}$, noting $\Gamma(n) = (n-1)!$ for integers n.

2.3 Dynamic Model

We have the dynamic component clonal model as

$$C_{ijk}\,|\,(\lambda_{ijk}, O_{jk}, D_{ij}=1) \sim \text{Pois}(\lambda_{ijk}O_{jk})$$

with

$$\lambda_{ijk}|(D_{ij}=1) \sim \text{Gamma}(\alpha, \beta)$$

For the dynamic model component we parameterize clonal mean λ_{ijk} to vary across time k. While λ_{ij} in the static model can be considered a random effect and has an expectation which is identifiable, λ_{ijk} is non-identifiable and and comparison of the two components is only possible under marginalization and the induced posterior predictive distributions. Integration of λ_{ijk} yields a product of posterior predictive distributions which are negative binomial distributed

$$\begin{aligned}
p(C_{ij\cdot}\,|\,O_{j\cdot}, D_{ij}=1, \alpha, \beta) &= \prod_{k=1}^{T_j} \int p(C_{ijk}, \lambda_{ijk}|O_{j\cdot}, D_{ij}=1, \alpha, \beta)\; d\lambda_{ijk} \\
&= \prod_{k=1}^{T_j} \int l(C_{ijk}|\lambda_{ijk},\, O_{j\cdot}) p(\lambda_{ijk}|D_{ij}=1, \alpha, \beta)\, d\lambda_{ijk} \\
&= \prod_{k=1}^{T_j} \int \frac{\exp(-\lambda_{ijk}\, O_{jk}) \lambda_{ijk}^{C_{ijk}} O_{jk}^{C_{ijk}}}{C_{ijk}!} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda_{ijk}^{\alpha-1} \exp\left(-\beta\lambda_{ijk}\right) d\lambda_{ijk} \\
&= \prod_{k=1}^{T_j} \frac{\Gamma(C_{ijk}+\alpha)}{\Gamma(\alpha) C_{ijk}!} \left(\frac{\beta}{\beta+O_{jk}}\right)^{\alpha} \left(\frac{O_{jk}}{\beta+O_{jk}}\right)^{C_{ijk}} \qquad (3)
\end{aligned}$$

where we write the distribution in terms of the $\Gamma(\cdot)$ function, and which in C_{ijk} we recognize as the product of T_j negative binomial mass functions on size parameter α and success probabilities $\beta/(\beta+O_{jk})$ for $k=1,\ldots,T_j$. We develop our mixture model using these mass functions in equations (2) and (3) for the static and dynamic components, respectively.

2.4 Comparing Components

To consider the circumstances under which a $C_{ij\cdot}$ vector will evaluate more favorably under the static or dynamic model, one can simplify their quotient and examine terms. Taking the static over that dynamic model, that is equation (2) over (3), yields a reduced expression for which static behavior would be associated with higher values:

$$\frac{p(C_{ij\cdot}\,|\,O_{jk}, D_{ij}=0, \alpha, \beta)}{p(C_{ij\cdot}\,|\,O_{jk}, D_{ij}=1, \alpha, \beta)} = \kappa_1 \cdot \frac{\Gamma(\sum_{k=1}^{T_j} C_{ijk} + \alpha)}{\prod_{k=1}^{T_j} \Gamma(C_{ijk}+\alpha)} \cdot \prod_{k=1}^{T_j} \left(\frac{\beta+O_{jk}}{\beta+\sum_k O_{jk}}\right)^{C_{ijk}} \qquad (4)$$

where the proportionality constant κ_1 is a function of α, β, and $O_{j\cdot}$, that is

$$\kappa_1 = \Gamma(\alpha)^{T_j-1} \left(\frac{\beta^{(1-T_j)} \prod_{k=1}^{T_j}(\beta+O_{jk})}{\beta+\sum_{k=1}^{T_j} O_{jk}}\right)^{\alpha}$$

We can multiply and divide by a ρ, and add and subtract $\alpha-1$ in the exponent and re-write the expression (4) as

$$= \kappa_1\kappa_2 \cdot \frac{\Gamma(\sum_{k=1}^{T_j} C_{ijk} + \alpha)}{\prod_{k=1}^{T_j} \Gamma(C_{ijk}+\alpha)} \cdot \prod_{k=1}^{T_j} \left(\frac{\rho(\beta+O_{jk})}{\beta+\sum_k O_{jk}}\right)^{C_{ijk}+\alpha-1} \qquad (5)$$

where ρ defined as follows allows the product terms to be interpreted as multinomial probabilities and κ_2 is a normalizing constant, with

$$\rho = \left[\sum_{k=1}^{T_j} \left(\frac{\beta+O_{jk}}{\beta+\sum_k O_{jk}}\right)\right]^{-1} \qquad \text{and} \qquad \kappa_2 = \rho^{-\sum_k (C_{ijk}+\alpha-1)} \prod_{k=1}^{T_j} \left(\frac{\beta+O_{jk}}{\beta+\sum_k O_{jk}}\right)^{-\alpha+1}$$

One observes that ρ is a function of β and $O_{j\cdot}$, and $\kappa_1\kappa_2$ is a function of $O_{j\cdot}, \sum_k C_{ijk}, \alpha, \beta$, and therefore constant for changes in $C_{ij\cdot}$ holding $\sum_k C_{ijk}$ constant.

If we conceptualize the expression (5) as conditioning on the $O_{j\cdot}$ and $\sum_k C_{ijk}$, this is multinomial and will evaluate highest with C_{ijk} as roughly proportional to $\beta+O_{jk}$ then offset by α. Assuming O_{jk} dominates β and C_{ijk} dominates α as is roughly the case in practice with low expected frequencies relative to high variation, the mode will be achieved when C_{ijk} is approximately proportion to O_{jk}. This is to say that C_{ijk}/O_{jk} is approximately constant and hence the quotient evaluates to a relatively high value up to the normalizing constant meaning the static component is favorable as compared to the dynamic component. In contrast, the expression will be minimized when in the tails of the multinomial, that is, when vector $C_{ij\cdot}$ is far from proportionality with O_{jk}. But holding $\sum_k C_{ijk}$ constant, this entails some C_{ijk}'s to be small and some large, which is to say the C_{ijk} exhibit dynamic and highly variable behavior.

2.5 Mixing and Model fitting

Having derived the component expressions we can write the likelihood in terms of them and the mixing distribution described with binary D_{ij} and $P(D_{ij} = 1|\pi) = \pi$. We write

$$\begin{aligned} p(C_{ij\cdot}\,|\,O_{j\cdot}.\alpha, \beta, \pi) &= p(C_{ij\cdot}\,, D_{ij} = 1|\,O_{j\cdot}, \alpha, \beta, \pi) + p(C_{ij\cdot}\,, D_{ij} = 0|\,O_{j\cdot}, \alpha, \beta, \pi) \\ &= p(C_{ij\cdot}\,|\,O_{j\cdot}, D_{ij} = 1, \alpha, \beta)P(D_{ij} = 1|\pi) + p(C_{ij\cdot}\,|\,O_{j\cdot}, D_{ij} = 0, \alpha, \beta)P(D_{ij} = 0|\pi) \\ &= \pi \cdot p(C_{ij\cdot}\,|\,O_{j\cdot}, D_{ij} = 1, \alpha, \beta) + (1 - \pi) \cdot p(C_{ij\cdot}\,|\,O_{j\cdot}, D_{ij} = 0, \alpha, \beta) \end{aligned}$$

since $P(D_{ij})$ is not a function of α nor β. Hyperparameters α and β are estimated using empirical Bayes and the model is fit using the EM algorithm [10]. We calculate the expectation of the complete data log-likelihood

$$\begin{aligned} \sum_{ijk} \log p(C_{ijk}|O_{jk}, \alpha, \beta, \pi) &= \sum_{ijk} \log \big(p(C_{ijk}, D_{ij} = 1|O_{jk}, \alpha, \beta, \pi) \\ &\qquad\qquad + p(C_{ijk}, D_{ij} = 0|O_{jk}, \alpha, \beta, \pi) \big) \\ &= \sum_{ijk} \log \big(\pi \cdot p(C_{ijk}|O_{jk}, D_{ij} = 1, \alpha, \beta) \\ &\qquad\qquad + (1 - \pi) \cdot\ p(C_{ijk}|O_{jk}, D_{ij} = 0, \alpha, \beta) \big) \end{aligned}$$

for the $(m + 1)^{st}$ iteration under $p(D_{ij}|C_{\cdot j\cdot}, O_{j\cdot}, \alpha^{(m)}, \beta^{(m)}, \pi^{(m)})$ given current estimates of $(\alpha^{(m)}, \beta^{(m)}, \pi^{(m)})$, then maximize the expression with respect to α, β, π to find $\alpha^{(m+1)}, \beta^{(m+1)}$, and $\pi^{(m+1)}$. That is, we find

$$Q(\alpha, \beta, \pi|\alpha^{(m)}, \beta^{(m)}, \pi^{(m)}) = E_{p_D}[\log\, p(C_{ijk}, D_{ij}|O_{jk}, \alpha, \beta, \pi)]$$

where for brevity of notation we use $p_D = p(D_{ij}|C_{ijk}, O_{jk}, \alpha^{(m)}, \beta^{(m)}, \pi^{(m)})$, and subsequently calculate

$$(\alpha^{(m+1)},\ \beta^{(m+1)},\ \pi^{(m+1)}) = \arg\max_{\alpha,\beta,\pi} Q(\alpha, \beta, \pi|\alpha^{(m)}, \beta^{(m)}, \pi^{(m)})$$

Since α and β are present in both components, maximization is difficult. While an additional, nested EM is possible to estimate $(\alpha^{(m+1)}, \beta^{(m+1)})$, here we use BFGS for the maximization step [4, 12].

Because our mixture consists of two components so that D_{ij} can be considered Bernoulli for all i, j, calculation of $p(D_{ij}|C_{ijk}, O_{jk}, \alpha^{(m)}, \beta^{(m)}, \pi^{(m)})$ is reduced to finding its expectation which is

$$\begin{aligned} E[D_{ij}|C_{ij\cdot}, O_{j\cdot},\alpha^{(m)}, \beta^{(m)}, \pi^{(m)}] &= p(D_{ij} = 1|C_{ij\cdot}, O_{j\cdot}, \alpha^{(m)}, \beta^{(m)}, \pi^{(m)}) \\ &= \frac{p(D_{ij} = 1|C_{ij\cdot}, O_{j\cdot}, \alpha^{(m)}, \beta^{(m)}, \pi^{(m)})}{p(D_{ij} = 1|C_{ij\cdot}, O_{j\cdot}, \alpha^{(m)}, \beta^{(m)}, \pi^{(m)}) + p(D_{ij} = 0|C_{ij\cdot}, O_{j\cdot}, \alpha^{(m)}, \beta^{(m)}, \pi^{(m)})} \\ &= \frac{p(D_{ij} = 1, C_{ij\cdot}|O_{j\cdot}, \alpha^{(m)}, \beta^{(m)}, \pi^{(m)})}{p(D_{ij} = 1, C_{ij\cdot}|O_{j\cdot}, \alpha^{(m)}, \beta^{(m)}, \pi^{(m)}) + p(D_{ij} = 0, C_{ij\cdot}|O_{j\cdot}, \alpha^{(m)}, \beta^{(m)}, \pi^{(m)})} \qquad (6) \\ &= \frac{\pi^{(m)}\ p(C_{ij\cdot}|O_{j\cdot}, D_{ij} = 1, \alpha^{(m)}, \beta^{(m)})}{\pi^{(m)}\ p(C_{ij\cdot}|O_{j\cdot}, D_{ij} = 1, \alpha^{(m)}, \beta^{(m)}) + (1 - \pi^{(m)})\ p(C_{ij\cdot}|O_{j\cdot}, D_{ij} = 0, \alpha^{(m)}, \beta^{(m)})} \qquad (7) \end{aligned}$$

where equation (6) uses $p(D_{ij}|C_{ij\cdot}, O_{j\cdot}, \alpha^{(m)}, \beta^{(m)}, \pi^{(m)}) \propto \pi^{(m)D_{ij}}(1{-}\pi^{(m)})^{1-D_{ij}}\ p(C_{ij\cdot}|O_{j\cdot}, D_{ij}, \alpha^{(m)}, \beta^{(m)})$, and also $p(D_{ij} = 1|C_{ij\cdot}, O_{j\cdot}, \alpha, \beta, \pi) = 1 - p(D_{ij} = 0|C_{ij\cdot}, O_{j\cdot}, \alpha, \beta, \pi)$, and (7) is calculated with component expressions (2) and (3). We iterate EM steps until the convergence criterion is reached, the mean of changes $\sum_{ij}(D_{ij}^{(m)} - D_{ij}^{(m+1)})/n$ for iteration m to $m + 1$ less than some ϵ, chosen here 10^{-8}.

3 Simulation

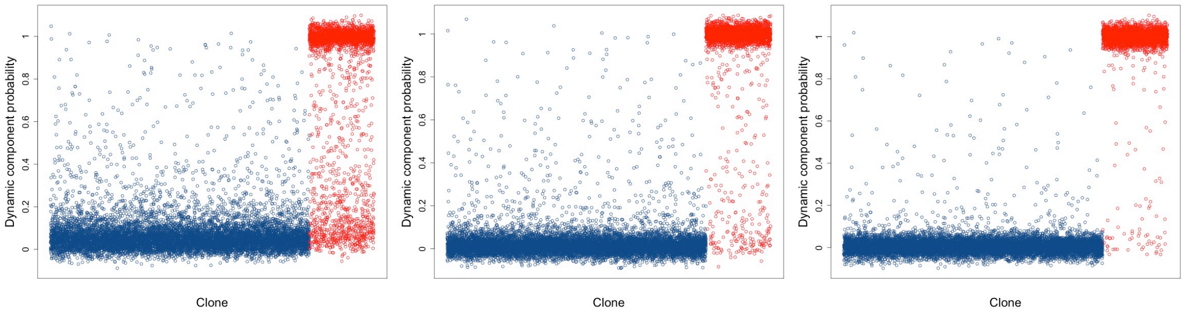
3.1 Generating process

We generated 60,000 clonotypes under a combination of α and β values as shown in Table 1 and with total observation-time offsets sampled from an exponential distribution that approximated the empirically observed total observation-time reads. Clonotype mean proportions were sampled from a Gamma(α, β) for the different parameter combinations, and subsequently multiplied by the offset O_{jk} to be used as the mean parameter in sampling from a Poisson distribution. This mean parameter was either varying or constant across the assumed four followup times as a function of whether the clone was considered dynamic or static, respectively. Depending on the simulation scenario, we fit the variance component mixture model to two, three, or four of the time points generated to examine the influence of followup on clonotype classification sensitivity and specificity and generated 20% of clones as dynamic and 80% static. We iterated 100 times for each parameter combination to assess variability. Parameter estimates for $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\pi}$ were calculated with four followup periods.

3.2 Simulation Results

Simulations indicate model fits reflect the parameters under which data are generated. Parameter estimates demonstrate a lack of bias and low variance as shown in Table 1. Convergence criteria under the parameter combinations described were achieved in fewer than 25 EM iterations.

Table 2 shows that under the parameters considered, specificity was nearly perfect and relatively invariant to the component membership threshold range of 0.75 and 0.95. Sensitivity to identifying dynamic clones decreased as a function of the threshold used, whereas it increased significantly in the number of followups, each holding the other variable constant and as expected. The results indicate that the more information one has within clone longitudinally, the greater the ability to distinguish between dynamic and static behavior. This behavior is demonstrated visually in Figure 1, as one can see more dense dynamic and static clones near the top and bottom, respectively, of the figures as the number of modelled followups increases suggesting less misclassification under any threshold.



(a) Clonotype membership probability with two observed follow-ups.

(b) Clonotype membership probability with three observed follow-ups.

(c) Clonotype membership probability with four observed follow-ups.

Figure 1: Model estimated clonotype membership probability for those generated under the static (blue points) versus dynamic models (red points) with two, three, and four within-clone observed follow-ups. Gaussian jitter is added with respect to the y-axis to decrease overplotting.

Scenario:					
α	β	π	$\hat{\alpha}$ (95% CI)	$\hat{\beta}$ (95% CI)	$\hat{\pi}$ (95% CI)
1	100	0.2	1.01 (1.01,1.02)	101.25 (100.32,102.41)	0.2 (0.2,0.2)
1	200	0.2	1.02 (1.02,1.03)	204.04 (202.11,206.48)	0.21 (0.2,0.21)
2	100	0.2	2.01 (1.99,2.02)	100.34 (99.28,101.33)	0.2 (0.2,0.2)
2	200	0.2	2.02 (2,2.03)	200.84 (199.19,203.23)	0.2 (0.2,0.2)

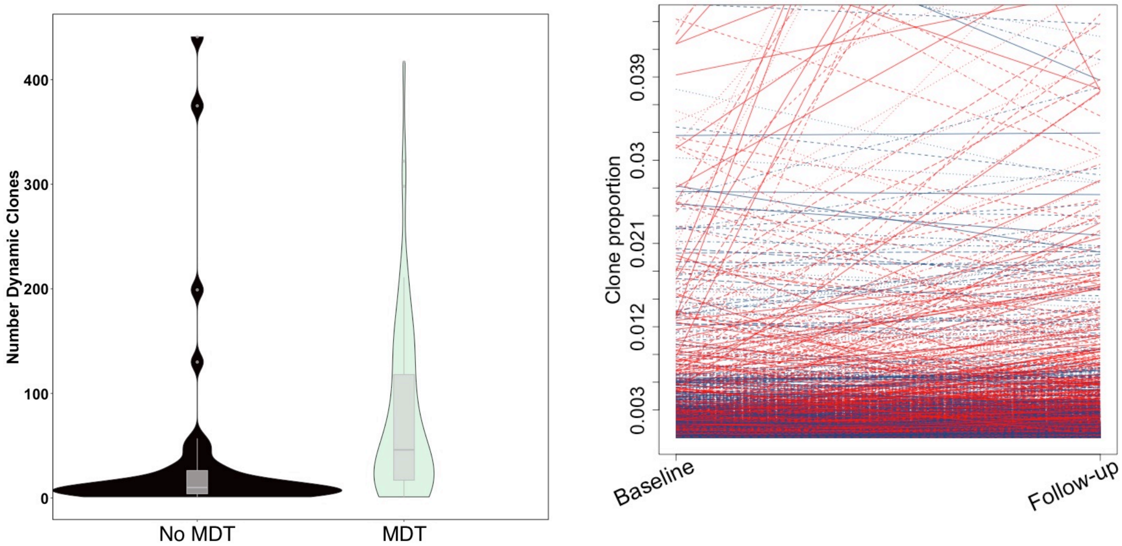
Table 1: Component and mixing parameter estimates and confidence intervals under four combinations of α and β based on 100 iterations of each scenario.

	Sensitivity		1-Specificity	
Membership Threshold:	0.75	0.95	0.75	0.95
2 observed times	0.70 (0.69,0.70)	0.64 (0.64,0.65)	0.99 (0.99,0.99)	1.00 (1.00,1.00)
3 observed times	0.90 (0.90,0.91)	0.88 (0.87,0.89)	1.00 (1.00,1.00)	1.00 (1.00,1.00)
4 observed times	0.97 (0.97,0.97)	0.96 (0.96,0.96)	1.00 (1.00,1.00)	1.00 (1.00,1.00)

Table 2: Sensitivity and specificity rates as a function of component membership threshold and number of observed followup period.

4 Results

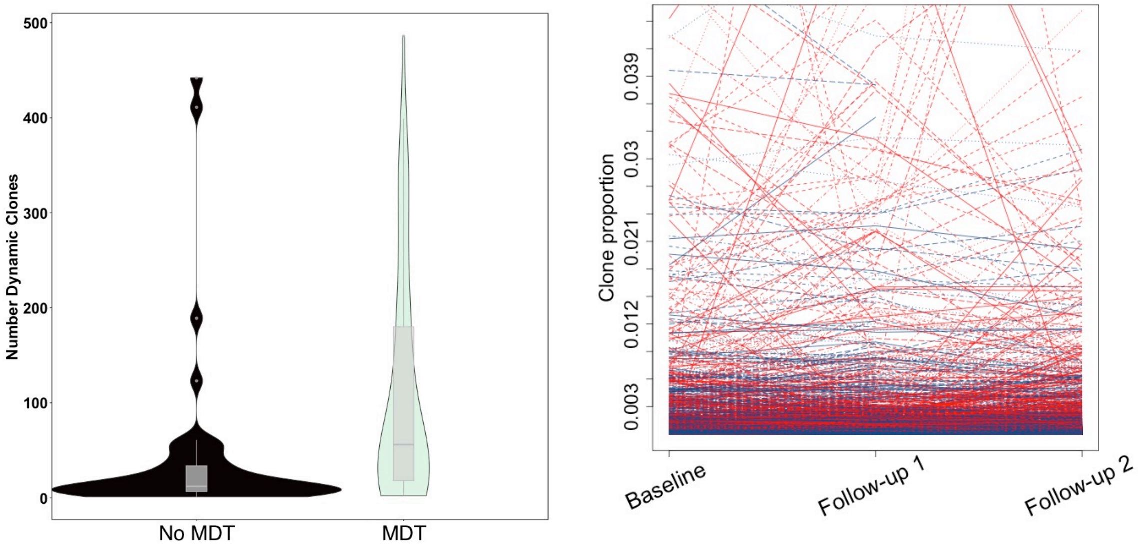
We fit the model to a longitudinal cohort of 108 prostate cancer patients to whom androgen-deprivation therapy was applied and metastasis-directed radiation therapy (MDT) was randomized [24, 16]. We performed analyses of baseline-followup cross-sections, and baseline with two followups, the latter including approximately 15% missingness across the second two followup periods. The total number of productive clonotypes considered across person-times was approximately 4.9 million. After filtering clones to have at least 8 template reads in total across baseline and followup periods, we arrived at 62,662 clones on 97 observations for the baseline-followup analysis, and 86,381 on 104 observations for the baseline with two followups analysis. For this latter analysis, one cross-section of followup occurred at the end of radiation for the MDT stratum and at the progression cross-section for the non-MDT stratum among patients who experienced disease progression. While only a subset of the non-MDT stratum experienced progression, one might expect it to increase model-defined dynamism because of immune response, thus pulling the non-MDT stratum toward MDT and therefore the null hypothesis with respect to that metric. For each person-time, the total number of template reads O_{jk} was calculated for use as the logged offset so that variation and therefore component membership would be interpreted on the scale of clonotype proportion, rather than the count of template reads. We fit the model to these data across MDT treatment strata, and also within treatment strata to examine heterogeneity in estimated $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\pi}$. Once the convergence criterion was reached, we examined cluster membership and parameter estimates. Using a threshold component membership probability of > 0.75 to determine the number of unique dynamic clones, we counted the number within patient and associated the measure with treatment stratum. For the baseline-followup analysis, we additionally partitioned the number of dynamic clones into those expanding and contracting based on the trajectory of the proportion change over baseline and followup.



(a) Densities of the number of dynamic clones with boxplot overlay, stratified by metastasis-directed radiation therapy.

(b) Proportion clonotype over time with dynamic clones shown in red and static clones in dark blue as estimated by the model. Static clones are downsampled by a factor of 2 for visibility of dynamic clones.

Figure 2: Two followup analysis of baseline-followup cross sections using variance component mixture model.



(a) Densities of the number of dynamic clones with boxplot overlay, stratified by metastasis-directed radiation therapy.

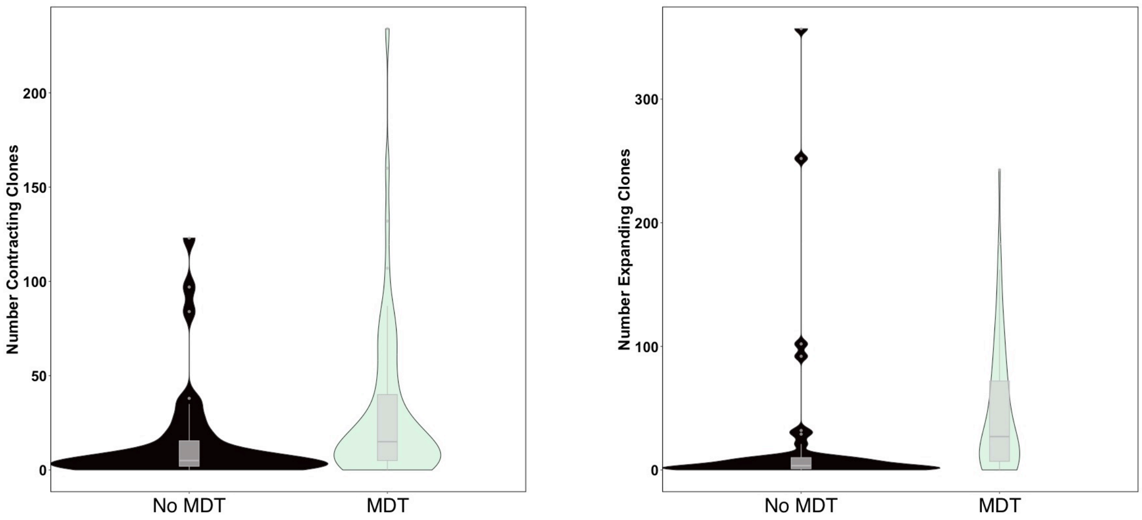
(b) Proportion clonotype over time with dynamic clones shown in red and static clones in dark blue as estimated by the model. Static clones are down-sampled by a factor of 2 for visibility of dynamic clones.

Figure 3: Three followup analysis using variance component mixture model.

Results indicate a significant increase in the number of dynamic clones under MDT for both longitudinal analyses. Figures 2a), 3a), and 4 all show qualitative marked increases in the number of dynamic, expanding, and contracting clones in the MDT treatment stratum, which are formally tested and found significant in nearly all cases by χ^2 tests and log-linear models (Table 3). The non-MDT stratum had a small number of outlier observations that exhibited a large number of dynamic clones, likely pulling test statistics more toward the null hypothesis of no treatment strata differences than would have otherwise been observed.

One observes in Figures 2b) and 3b) that dynamic clones tend to exhibit significant slopes as the clonotype proportion changes dramatically across followup, in contrast to static clones which are flatter. One also sees that the higher average clonotype proportion, the larger absolute change in proportion is needed for a clone to be considered dynamic, incorporating expected biological variability into the model in assessing dynamic behavior.

Within MDT treatment strata, models reveal some heterogeneity in estimated α and β as shown in Table 4, suggesting differences in shape of clonal dynamics over time as a function of treatment. Since α and β govern both the dynamic and static mixture components, interpretation is not straightforward, though the implication of estimates is that among the MDT stratum, clonotype proportions tend to be smaller and have smaller variance than the non-MDT stratum across time and components. Since clonotype proportions are sensitive to the number of unique clones within observation-times because they must sum to one, interpretation might indicate a tendency toward greater "focus" in dynamism among MDT stratum clones as those useful in immune response expand and those that are not contract.



(a) Contracting clone count densities with boxplot overlay, stratified by metastasis-directed radiation therapy.

(b) Expanding clone count densities with boxplot overlay, stratified by metastasis-directed radiation therapy.

Figure 4: Analysis of baseline-followup cross sections using variance component mixture model for number of contractions and number of expansions.

	2 observed follow-ups (FU's)			3 FU's
Clone group:	Dynamic	Expansions	Contractions	Dynamic
χ^2 test p-value	0.0007	0.0003	0.13	0.001
log-linear parameter (p-value)	0.68 (10^{-8})	0.66 (10^{-8})	0.71 (10^{-8})	0.98 (10^{-8})

Table 3: Hypothesis tests for dynamic clone stratification by MDT treatment stratum. χ^2 tests are based on dichotomizing at 50 total dynamic clones or 25 total expanding/contracting clones depending on the case. Log-linear model parameter indicates the log expected proportion increase in the number of dynamic, expanding, or contracting clones, depending on the case, for the MDT stratum as compared to no MDT.

	2 observed follow-ups			3 observed follow-ups		
	Combined	No MDT	MDT	Combined	No MDT	MDT
$\hat{\alpha}$	0.35	0.36	0.34	0.33	0.37	0.32
$\hat{\beta}$	640.00	603.41	658.44	715.33	648.81	752.05
$\hat{\pi}$	0.23	0.22	0.24	0.20	0.19	0.20

Table 4: Model parameter estimates for two and three study follow-ups, across and within the metastasis-directed radiation therapy (MDT) and lack of MDT (No MDT) treatment strata.

5 Discussion

In this work we placed T-cell receptor clonal dynamics in a mixture model framework grounded in Bayesian hierarchy and argued the approach can be a useful tool for understanding these high-dimensional data that duly acknowledges expected biological variability while incorporating fundamental model features like variable longitudinal followup and missingness. Since receptors shared across subjects are rare relative to those unique to each subject, traditional statistical high-dimensional regression tools are not as useful and one must try to relate summarizations of clonal population dynamics to subject phenotype, in our case a count of model-determined dynamic clones. Additionally partitioning dynamic clones into those that expand and contract as a function of their trajectory further allows one to examine the kind of dynamism most sensitive to intervention, in our case metastasis-directed radiation therapy. Early identification of clinically relevant dynamic clones shortly following therapy, such as metastasis-directed radiation therapy, may facilitate treatment decision-making and subsequent therapeutic approaches, representing a promising actionable biomarker underlying host immunobiology. In addition, resolution of the systemic TCR repertoire through this approach is more practical and feasible than serial tissue biopsies, particularly as co-culturing assays of T cells with tumor tissue are laborious, resource intensive, and can be challenging to scale.

The mixture of negative multinomial and product of negative binomial components arrived at through the posterior predictive distribution of Gamma-Poisson conjugacy was both theoretically convenient and reflective of observed data and therefore a promising means of modelling these data. However, future work may well attempt flexible alternatives for which more computationally intensive fitting procedures are necessary. Since the single parameter Poisson distribution equates the mean and variance, enforcing the relationship between clone frequency and degree of variability constituting “dynamism”, a model allowing one to specify that relationship could help clinicians and subject experts place greater weight on, say, dynamism among the low-frequency group of clones or vice versa, depending on hypothesized biological mechanism. The mean-variance equivalence of the Poisson distribution does have advantages in that since coverage increases in the root of the variance, one is better powered to identify dynamism among higher frequency clones for less variability relative to that mean, consistent with a biological prior that higher frequency clones are inherently more interesting and should be easier to label as dynamic. Still, exploring sensitivity to different and more flexible mean-variance specifications could improve characterization of the dynamism associated with subject phenotype.

References

[1] Murray Aitkin. “A general maximum likelihood analysis of variance components in generalized linear models”. In: *Biometrics* 55.1 (1999). Publisher: Wiley Online Library, pp. 117–128.

[2] Laura R. E. Becher et al. “Public and private human T-cell clones respond differentially to HCMV antigen when boosted by CD3 copotentiation”. en. In: *Blood Advances* 4.21 (Nov. 2020), pp. 5343–5356. ISSN: 2473-9529, 2473-9537. DOI: 10.1182/bloodadvances.2020002255. URL: https://ashpublications.org/bloodadvances/article/4/21/5343/469790/Public-and-private-human-Tcell-clones-respond (visited on 11/29/2023).

[3] Yoav Benjamini and Yosef Hochberg. “Controlling the false discovery rate: a practical and powerful approach to multiple testing”. In: *Journal of the Royal statistical society: series B (Methodological)* 57.1 (1995). Publisher: Wiley Online Library, pp. 289–300.

[4] Charles George Broyden. “The convergence of a class of double-rank minimization algorithms 1. general considerations”. In: *IMA Journal of Applied Mathematics* 6.1 (1970). Publisher: Oxford University Press, pp. 76–90.

[5] Gilles Celeux, Sylvia Frühwirth-Schnatter, and Christian P. Robert. “Model Selection for Mixture Models – Perspectives and Strategies”. en. In: *Handbook of Mixture Analysis.* Ed. by Sylvia Frühwirth-Schnatter, Gilles Celeux, and Christian P. Robert. 1st ed. Boca Raton, Florida : CRC Press, [2019]: Chapman and Hall/CRC, Jan. 2019, pp. 117–154. ISBN: 978-0-429-05591-1. DOI: 10.1201/9780429055911-7. URL: https://www.taylorfrancis.com/books/9780429508240/chapters/10.1201/9780429055911-7 (visited on 03/03/2023).

[6] Gilles Celeux, Sylvia Früwirth-Schnatter, and Christian P Robert. “Handbook of Mixture Analysis”. en. In: ().

[7] Ziyi Chen et al. “Decreased Treg Cell and TCR Expansion Are Involved in Long-Lasting Graves’ Disease”. en. In: *Frontiers in Endocrinology* 12 (Apr. 2021), p. 632492. ISSN: 1664-2392. DOI: 10.3389/fendo.2021.632492. URL: https://www.frontiersin.org/articles/10.3389/fendo.2021.632492/full (visited on 11/29/2023).

[8] Shin-Heng Chiou et al. “Global analysis of shared T cell specificities in human non-small cell lung cancer enables HLA inference and antigen discovery”. en. In: *Immunity* 54.3 (Mar. 2021), 586–602.e8. ISSN: 10747613. DOI: 10.1016/j.immuni.2021.02.014. URL: https://linkinghub.elsevier.com/retrieve/pii/S1074761321000807 (visited on 11/29/2023).

[9] Paul Delmar et al. “Mixture Model on the Variance for the Differential Analysis of Gene Expression Data”. en. In: *Journal of the Royal Statistical Society Series C: Applied Statistics* 54.1 (Jan. 2005), pp. 31–50. ISSN: 0035-9254, 1467-9876. DOI: 10.1111/j.1467-9876.2005.00468.x. URL: https://academic.oup.com/jrsssc/article/54/1/31/7113006 (visited on 07/11/2024).

[10] Arthur P Dempster, Nan M Laird, and Donald B Rubin. “Maximum likelihood from incomplete data via the EM algorithm”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1 (1977). Publisher: Wiley Online Library, pp. 1–22.

[11] William S DeWitt et al. “Dynamics of the cytotoxic T cell response to a model of acute viral infection”. In: *Journal of virology* 89.8 (2015). Publisher: Am Soc Microbiol, pp. 4517–4526.

[12] Roger Fletcher. *Practical methods of optimization.* John Wiley & Sons, 2000.

[13] Sylvia Frühwirth-Schnatter and Sylvia Frèuhwirth-Schnatter. *Finite mixture and Markov switching models.* Vol. 425. Springer, 2006.

[14] Isobel Claire Gormley and Sylvia Frühwirth-Schnatter. *Mixtures of Experts Models.* en. arXiv:1806.08200 [stat]. June 2018. URL: http://arxiv.org/abs/1806.08200 (visited on 08/19/2022).

[15] Sara Hey et al. “Analysis of CDR3 Sequences from T-Cell Receptor Beta in Acute Respiratory Distress Syndrome”. en. In: *Biomolecules* 13.5 (May 2023), p. 825. ISSN: 2218-273X. DOI: 10.3390/biom13050825. URL: https://www.mdpi.com/2218-273X/13/5/825 (visited on 11/29/2023).

[16] Ethan B Ludmir et al. "Addition of metastasis-directed therapy to systemic therapy for oligometastatic pancreatic ductal adenocarcinoma (EXTEND): A multicenter, randomized phase II trial". In: *Journal of Clinical Oncology* 42.32 (2024), pp. 3795–3805.

[17] Huaichao Luo et al. "Characteristics and significance of peripheral blood T-cell receptor repertoire features in patients with indeterminate lung nodules". en. In: *Signal Transduction and Targeted Therapy* 7.1 (Oct. 2022), p. 348. ISSN: 2059-3635. DOI: 10.1038/s41392-022-01169-7. URL: https://www.nature.com/articles/s41392-022-01169-7 (visited on 11/29/2023).

[18] Kevin P. Murphy. *Machine learning: a probabilistic perspective*. en. Adaptive computation and machine learning series. Cambridge, MA: MIT Press, 2012. ISBN: 978-0-262-01802-9.

[19] Dechavudh Nityasuddhi and Dankmar Böhning. "Asymptotic properties of the EM algorithm estimate for normal mixture models with component specific variances". In: *Computational statistics & data analysis* 41.3-4 (2003). Publisher: Elsevier, pp. 591–601.

[20] Christian P Robert, George Casella, and George Casella. *Monte Carlo statistical methods*. Vol. 2. Springer, 1999.

[21] Elisa Rosati et al. "Overview of methodologies for T-cell receptor repertoire analysis". en. In: *BMC Biotechnology* 17.1 (Dec. 2017), p. 61. ISSN: 1472-6750. DOI: 10.1186/s12896-017-0379-9. URL: http://bmcbiotechnol.biomedcentral.com/articles/10.1186/s12896-017-0379-9 (visited on 11/29/2023).

[22] Julie Rytlewski et al. "Model to improve specificity for identification of clinically-relevant expanded T cells in peripheral blood". en. In: *PLOS ONE* 14.3 (Mar. 2019). Ed. by Paul A. Beavis, e0213684. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0213684. URL: https://dx.plos.org/10.1371/journal.pone.0213684 (visited on 11/29/2023).

[23] Shayle R Searle, George Casella, and Charles E McCulloch. *Variance components*. John Wiley & Sons, 2009.

[24] Chad Tang et al. "Addition of Metastasis-Directed Therapy to Intermittent Hormone Therapy for Oligometastatic Prostate Cancer: The EXTEND Phase 2 Randomized Clinical Trial". en. In: *JAMA Oncology* 9.6 (June 2023), p. 825. ISSN: 2374-2437. DOI: 10.1001/jamaoncol.2023.0161. URL: https://jamanetwork.com/journals/jamaoncology/fullarticle/2803085 (visited on 07/06/2023).

[25] Yvonne H. F. Teng et al. "Analysis of T cell receptor clonotypes in tumor microenvironment identifies shared cancer-type-specific signatures". en. In: *Cancer Immunology, Immunotherapy* 71.4 (Apr. 2022), pp. 989–998. ISSN: 0340-7004, 1432-0851. DOI: 10.1007/s00262-021-03047-7. URL: https://link.springer.com/10.1007/s00262-021-03047-7 (visited on 11/29/2023).

[26] Ashenafi A Yirga et al. "Negative binomial mixed models for analyzing longitudinal CD4 count data". In: *Scientific reports* 10.1 (2020). Publisher: Nature Publishing Group UK London, p. 16742.