arXiv:2601.04536v1 [q-bio.QM] 8 Jan 2026

利用纵向贝叶斯混合模型识别扩增的TCR克隆型及其与癌症患者预后、转移灶定向治疗和VJ基因富集的关联

David Swanson*[1]、Alexander Sherry[5], 、Cara Haymaker[2], 、Alexandre Reuben[3], 和Chad Tang[4]

[1]生物统计学系，[2]转化分子病理学系，[3]胸部/头颈部内科肿瘤学系，[4]德克萨斯大学MD安德森癌症中心放射肿瘤学系；[5]梅奥诊所放射肿瘤学系

2025年12月

摘要

近年来，研究T细胞受体（TCR）的克隆性已成为了解免疫系统对癌症及其相关干预措施反应的一种方式。这类分析的一个方面是确定哪些受体会因治疗干预等外源性扰动而发生具有统计学显著性的扩增或收缩。我们对这类分析中常用的Fisher精确检验方法进行了阐述，并提出一种无需在患者内进行两两比较的替代方法。我们构建了这种灵活的贝叶斯纵向混合模型，可适应不同长度的患者随访，并能处理存在的缺失数据；即使出于实际操作上的限制，也不会在估计时舍弃数据。模型将克隆划分为动态类别（扩增或收缩）和静态类别后，可将其计数或其他特征与疾病状态、干预措施、基线生物标志物及患者预后关联起来。我们将这些方法应用于一组前列腺癌患者队列，这些患者接受了随机分配的转移灶定向治疗或未接受该治疗。分析发现，MDT患者的克隆扩增显著增加，且无论是否按MDT分层，这种增加均与后续疾病进展相关。我们还利用所构建的高维惩罚对数线性模型分析受体基序与VJ基因富集组合，结果提示扩增克隆具有独特的生物学特征，无论其扩增是否由MDT诱导。

关键词 ： T细胞受体、VDJ基因、基序分析、贝叶斯混合模型、惩罚对数线性模型 el

1 引言

适应性免疫系统通过识别并响应体内存在的病原体，在一定程度上对人类健康发挥着核心作用。T细胞是其中的关键组成部分，其同名受体可通过非同源重组实现极高的多样性；据估计，可产生多达 10^{18} 种不同的T细胞受体（TCR）基因 [15]。近年来，随着相关技术的发展，以及人们对其与免疫功能之间关联的研究和认识不断加深，对TCR库的分析日益普遍 [29, 16, 2, 5]。

尽管T细胞受体的潜在多样性和实际多样性都极为丰富，但只有其中一部分会在免疫反应过程中表现出扩增或收缩行为。由于统计噪声是任何数据环境中的核心问题，因此，判定哪些克隆属于扩增、收缩，或更一般地说属于动态克隆（当存在多于2个纵向截面时），既是一个普遍重要的分析问题，也关系到癌症患者的预后 [6, 13]。本文的首项贡献是提出一种纵向贝叶斯

*ORCID：0000-0003-3174-1656

混合模型，该模型由两个组分构成：一个用于拟合扩增或收缩的纵向克隆频率，另一个用于拟合静态频率。通过使用哈密顿蒙特卡洛（HMC）在庞大的参数空间中进行采样来拟合模型后，便可以对特定克隆在基线-随访分析中是否表现出扩增、收缩或静态行为，或一般情况下是否表现为动态而非静态行为，作出概率性判断 [3, 12]。遵循纵向统计学的传统，在提出该模型时，我们寻求比TCR分析中常用的Fisher精确检验和贝塔-二项方法更高的通用性 [11, 10, 24], ；后两者主要适用于成对且完全观测的情形，而我们的方法则能够适应随机缺失（MAR）机制和任意随访 [17, 20, 19]。我们还注意到，混合模型所提供并产生的关于归属于动态类别或静态类别之一的概率性判断，使得在将动态克隆与患者结局相关联时可以采用灵活的阈值设定——下游分析可能显示，高概率扩增克隆的数量与患者结局相关，而采用较低严格度阈值时得到的相应数量则不相关。

受体多样性的一个原因是其由V、D和J基因构成。文献中已有一些关于V、D和J基因家族相互作用或富集的分析，即考察共同出现的基因家族组合是否比根据边际总数所预期的比例更高或更低，但关于这一主题的大多数研究都采用可视化或描述性方法 [22, 26, 30, 7] 。因此，VDJ富集分析中一个亟待解决的重要问题是，如何计算跨样本、在接受特定干预的样本中，或结合上述建模框架，在表现出不同扩增和收缩行为的T细胞受体群体中出现的基因共现模式的统计显著性。

分类数据分析中一类成熟的模型，且特别适合评估VJ基因家族共同富集是否具有统计显著性，即对数线性模型 [1]。这类模型是若干模型类别（如逻辑回归）的推广，能够对任意形状和类别数的列联表中的相互作用进行灵活编码 $K_1 \times K_2 \times \cdots \times K_N$。通常，人们会依据专业知识提出假设并确定结构，进而进行对数线性建模；因此，VJ富集分析面临的挑战是需要一个无假设且灵活的模型空间。为此，在采用完全饱和模型参数化时，我们引入一个由数据估计的L1惩罚参数，通过将过于庞大且灵活的模型空间中的大量参数收缩至零，实现稀疏模型拟合 [28]。这一框架构成了本文的第二项方法学贡献。

我们首先在第2.1节和第2.2节中，通过分析并比较 Fisher精确检验方法与我们模型的误差模型，探讨如何区分静态克隆与扩增/收缩克隆。在第2.3节中，我们详细介绍所提出的贝叶斯混合模型，作为一种替代方案。最后，我们在第2.4节介绍用于识别 VJ基因与患者特征联合富集模式的对数线性L1惩罚回归模型框架。在第3节中，我们对一个前列腺癌队列进行了深入研究；该队列中的患者被随机分配接受转移灶定向治疗（MDT）或不接受该治疗（无MDT），且每组均包含接受额外间歇性或持续性雄激素剥夺治疗（ADT）的患者。我们运用方法部分开发的工具，将模型拟合到纵向T细胞受体频率数据，以识别扩增、收缩或通常意义上的动态克隆型数量。我们分析了这些组别及其数量与患者特征之间的关联，包括MDT、受体基序、基线生物标志物和无进展生存期（PFS）。此外，熵分析和分布分析显示，扩增克隆与收缩克隆之间存在轨迹差异。最后，应用惩罚对数线性模型框架发现，若干VJ基因家族的富集具有统计学显著性，并发现了一些相互作用及克隆型扩增状态。我们在第4节总结讨论分析所揭示的临床和生物学见解、潜在治疗靶点，以及改进贝叶斯混合模型和惩罚对数线性模型的可能方法。

2 方法

2.1 用于评估纵向克隆变化的 Fisher精确检验误差模型

我们首先考虑 Fisher精确检验的误差模型；在考察该检验有能力检测的扩增和收缩克隆轨迹时，这一模型将证明非常有用。考虑个体 j 在时间 k、 C_{ijk}的 i^{th} 克隆模板计数，以及个体 j 在时间k、 $O_{jk} = \sum_{i=1}^{U_{jk}} C_{ijk}$ 的模板计数总数，其中 U_{jk} 是个体 j 在时间 k的独特克隆数量。由于模板读数总数存在显著变异，而从生物学角度来看，克隆的比例如何

随时间发生变化是主要关注点，因此我们对比例（即频率） C_{ijk}/O_{jk} 随时间和患者进行建模。

在 Fisher精确法中，人们沿纵向随访选取这些测量值的成对数据，并对其变化进行假设检验 [9, 10] ；同样，也可以采用带有弱信息贝塔先验的变异形式 [24]。这些都是合理的方法，并且已被成功应用；我们可以尝试了解它们的统计特性。考虑这样一种检验：我们观测到C_{ij1} 和 C_{ij2}，即个体 j 在时间点1和2的克隆型 i 的模板计数，且不失一般性；相应地， O_{j1} 和 O_{j2}则是这两个时间点的总读数。

使用超几何分布检验以下零假设： C_{ijk} 与 O_{jk} 的比例在 $k = 1$，2期间保持不变。因此，将 $k = 1$，2的这4个单元格计数 C_{ijk} 和 $O_{jk} - C_{ijk}$ 设想为一个 2×2 表（表1），并在给定所有边际总数的条件下，可以计算 Fisher精确检验的 P值。

	时间 1	时间 2
克隆 ij	C_{ij1}	C_{ij2}
非克隆 ij	$O_{j1} - C_{ij1}$	$O_{j2} - C_{ij2}$

表 1： 2×2 个体 j在时间点 $k = 1$，2 的 TCR 模板读数与总读数列联表

众所周知，随着单元格计数增加，Fisher精确检验会越来越接近通常的比例相等评分检验，并收敛到某个比例 $C_{ijk}/O_{jk} \to p_k$ ，适用于 $k = 1$，2；两者的区别在于，前者仅对纵轴上的两个单元格边际分布进行条件化，而后者还对横轴上的边际分布进行条件化 [4]。为便于表述和阐释，我们因此考虑检验统计量 $\hat{}p_1 = C_{ij1}/O_{j1}$ 和 $\hat{}p_2 = C_{ij2}/O_{j2}$ 以及 $\hat{}p = (C_{ij1} + C_{ij2})/(O_{j1} + O_{j2})$

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{(1/O_{j1} + 1/O_{j2}) \cdot \hat{p}(1 - \hat{p})}} \sim N(0, 1) \text{ under } H_0$$

在零假设 [23] 下，该统计量服从标准正态分布。注意分母中的方差估量，其幂次取决于个体-时间点 jk 的模板读数总数，而且其形式为 $p(1 - p)$。对于较小的 p——低频克隆型即属此类——$1 - p \approx 1$，因此单位方差近似为 $p \approx p(1 - p)$。然而，对于频率较高的克隆型，则不能忽略 $1 - p$ 项，因此方差为 $O(p^2)$。由于 $p < 1$，这意味着方差更小，因而标准差也更小。在其他条件相同的情况下，由于检验统计量的分母较小，高频克隆只需较小的比例变化，就会被视为显著扩增或收缩。这一现象叠加在误差覆盖范围下降之上，因此，当均值随 λ 线性增长而标准差按平方根速率增长时，对于某个固定的纵向变化，检验效能也会提高。在某些情况下，这种做法可能是合适的，因为将高频克隆更受关注这一生物学先验纳入考量或许有其必要。然而，这种做法可能未能充分考虑生物学变异；人们可能希望将其设定为更接近克隆平均频率的比例变化，从而避免将微不足道的生物学变化标记为具有统计学显著性的动态行为——也就是说，即便频率从 0.5% 变为 1.0% 应被视为显著变化，也不希望把频率从 11.5% 变为 11% 这样的变化判定为显著变化。

2.2 替代误差模型

为了对误差项以及何种情况构成显著生物学变异保持一致的看法，对于某些 λ ，若其分布具有重要性质 $E[C_{ijk}] = Var[C_{ijk}] = \lambda \cdot O_{jk}$，似乎最好采用模型 $C_{ijk} \sim \text{Pois}(\lambda \cdot O_{jk})$，并将 O_{jk} 视为偏置项。请注意，由于随机变量的误差覆盖范围随标准差而非方差变化，这种表述仍然纳入了这样一种生物学先验：更关注频率较高的克隆，同时保持对误差项更为一致的看法。这意味着，与频率较低的克隆相比，对于频率较高的克隆中的比例变化，模型具有更高的检验效能。这是因为对于某个与均值成比例的变化，记作 $\delta \cdot E[C_{ijk}]$ ，对于某个 $0 < \delta < 1$，其变异系数的倒数（与检验效能相关的量）为

$$\frac{\delta \cdot E[C_{ijk}]}{\sqrt{Var[C_{ijk}]}} = \delta \frac{\lambda \cdot O_{jk}}{\sqrt{\lambda \cdot O_{jk}}} = \delta \sqrt{\lambda \cdot O_{jk}}$$

随着 λ（解释为期望频率）的表达值增大，并保持 δ 不变，随着平均频率增加，我们便能更有检验效能地检测微小变化。

2.3 拟议的模型层级

在 Fisher精确检验的设定中，零假设意味着各时间点 k 的成功概率 p_k 相同，因此使用上述符号，我们有 H_0： $p_1 = p_2$。与之类似，我们提出的模型旨在区分泊松强度参数 λ 是各时间点均相同还是随时间变化。因此，我们采用两种不同的参数化：一种是静态的 λ_{ij}，另一种则会变化，即随时间动态变化的 λ_{ijk}。也就是说，我们有

$$\text{Dynamic: } C_{ijk} \,|\, O_{jk} \sim \text{Pois}(\lambda_{ijk} O_{jk}) \quad \text{Static: } C_{ijk} \,|\, O_{jk} \sim \text{Pois}(\lambda_{ij} O_{jk}) \tag{1}$$

值得注意的是，虽然简单的 Fisher精确检验或比例检验会比较时间上的 2 个横截面，但此处我们用某个 $k = 1, \ldots..$ T_j进行一般化索引，其中 T_j 是观测值 j的随访时间数量，其中一些随访时间对于 C_{ijk}可能缺失，因此这是一种更一般的形式。由于 λ 只有一个自由度，边际分布不够灵活，无法充分描述观测到的数据，同时贝叶斯共轭性也为此提供了动机，因此我们提出以下层级形式：

$$\lambda_{ijk} \,|\, \alpha_j, \beta_j \sim \text{Gamma}(\alpha_j, \beta_j) \qquad \lambda_{ij} \,|\, \alpha_j, \beta_j \sim \text{Gamma}(\alpha_j, \beta_j) \tag{2}$$

注意到 α_j 和 β_j 在两个成分之间是共同的，并且用 j 对它们进行索引，可以处理由不同个体之间独特克隆数量的变异所引起的个体内相关性。

利用共轭关系，在对相应成分中的 λ_{ij} 或 λ_{ijk} 进行积分消去后，可得到以下简便的概率质量函数：

$$\begin{aligned} p_S(C_{ij\cdot} | O_{j\cdot}, \alpha_j, \beta_j) &= \int p_S(C_{ij\cdot}, \lambda_{ij} | O_{jk}, \alpha_j, \beta_j)\, d\lambda_{ij} = \int l_S(C_{i\cdot} | \lambda_{ij}\, O_{jk})\, p(\lambda_{ij} | \alpha_j, \beta_j)\, d\lambda_{ij} \\ &= \frac{\Gamma(\sum_k C_{ijk} + \alpha_j)}{\Gamma(\alpha_j) \prod_{k=1}^{T_j} C_{ijk}!} \left(\frac{\beta_j}{\beta_j + \sum_k O_{jk}} \right)^{\alpha_j} \prod_{k=1}^{T_j} \left(\frac{O_{jk}}{\beta_j + \sum_k O_{jk}} \right)^{C_{ijk}} \end{aligned} \tag{3}$$

$$\begin{aligned} p_D(C_{ij\cdot} | O_{j\cdot}, \alpha_j, \beta_j) &= \int p_D(C_{ij\cdot}, \lambda_{ijk} | O_{jk}, \alpha_j, \beta_j)\, d\lambda_{ijk} = \int l_D(C_{i\cdot} | \lambda_{ijk}\, O_{jk})\, p(\lambda_{ijk} | \alpha_j, \beta_j)\, d\lambda_{ijk} \\ &= \prod_{k=1}^{T_j} \frac{\Gamma(C_{ijk} + \alpha_j)}{\Gamma(\alpha_j) C_{ijk}!} \left(\frac{\beta_j}{\beta_j + O_{jk}} \right)^{\alpha_j} \left(\frac{O_{jk}}{\beta_j + O_{jk}} \right)^{C_{ijk}} \end{aligned} \tag{4}$$

动态成分和静态成分的混合形式为

$$p(C_{ij\cdot} | O_{j\cdot}, \alpha_j, \beta_j) = p_S(C_{ij\cdot} | O_{j\cdot}, \alpha_j, \beta_j) + \pi \cdot p_D(C_{ij\cdot} | O_{j\cdot}, \alpha_j, \beta_j)$$

其中，动态克隆的边际或总体混合比例为 π。我们在 A2 节中更详细地展示了推导过程。需要指出的是，分别对各自表达式中的 λ_{ij} 和 λ_{ijk} 进行积分消去后，可得到负多项分布或负二项分布乘积的概率质量函数，分别作为静态成分和动态成分的后验预测分布，并根据偏移项进行修正。按个体对 α_j 和 β_j 进行索引增加了灵活性，并允许患者体内具有可变数量的独特克隆；这是有益的，因为在对频率进行建模时，其总和等于 1 的线性依赖关系是固有的。

经验估计的 α_j 和 β_j 将在混合模型的“检验效能”中发挥调节作用，即模型区分落入相应成分的克隆型的能力。由于它们是基于实际观测到的频率进行经验估计的，因此不受分析人员所选择的偏差影响。在继续考察这两个成分的概率质量函数时，值得注意的是，对于其估计的 α_j 和 β_j 意味着较小均值和方差的数据，区分静态行为与动态行为的检验效能会较低，恰恰是因为上述原因——当从不同的 $\lambda_{ijk} \,|\, \alpha_j$， β_j ~Gamma(α_j, β_j) 中抽取样本，且其均值和方差较小时，泊松误差会压过层级结构的下一级变异性，因而无法区分这两个成分。事实上，均值可以任意大而方差很小，但趋势以及最坏情况下的情形将是均值小且方差小，因为随着均值增大，泊松误差的覆盖范围相对于均值会变小。

为了在总体间共享信息，我们引入另一个层级：

$$\log \alpha_j \sim N(\mu_\alpha, \sigma_\alpha^2) \qquad \log \beta_j \sim N(\mu_\beta, \sigma_\beta^2) \tag{5}$$

其中， μ_α、 μ_β、 σ_α^2和 σ_β^2 将通过经验估计，并对参数应用对数或逻辑斯蒂变换，以使其关于零对称。我们对 π 设置弥散的弱先验，断言

$$\text{logit}\, \pi \sim N(0, \sigma_\pi^2)$$

对于较大的 σ_π^2 ，并使用在STAN中实现的哈密顿蒙特卡洛（HMC）拟合模型，这是由于其层级形式以及缺乏闭式积分 [14, 27]。拟合过程对参数的恢复在A1节中进行了展示。

2.4 用于VJ基因富集检验及与患者表型相互作用的惩罚对数线性建模

我们以提出一种不同的模型来结束方法部分。该模型与上述用于将扩增和收缩克隆同静止克隆区分开来的模型具有互补的用途。TCR库分析的一个方面是考察V、D和J基因重组，以及这些基因家族单独或组合时是否存在基因家族富集。据我们所知，VDJ分析尚无用于对这种富集进行建模并检验其统计显著性的正式框架。在此，我们提出使用L1惩罚和完全饱和对数线性模型，对基因家族共现的列联表进行参数化，以此描述这些数据。采用这一框架时，还可以引入T细胞受体的其他特征，例如其是否发生扩增或收缩，从而检验基因家族与该特征之间的关联。在下文的模型描述中，为便于阐述并与第3节的分析相衔接，我们的记号仅聚焦于 V 和 J 基因，以及用 P表示的某种任意TCR或患者特征。

令 G_{pvj} 为具有患者特征 p、 V V基因家族 v 和 J J基因家族 j 的克隆型数量。也就是说， G_{pvj} 是由这些基因家族组成、且在表型为 p 的患者 j 中发现的 $C_{ij\cdot}$ 数量。然后定义

$$f(\beta_p^P, \ldots, \beta_{pvj}^{PVJ}) = f(\beta_p^P, \beta_v^V, \beta_j^J, \beta_{pv}^{PV}, \beta_{pj}^{PJ}, \beta_{vj}^{VJ}, \beta_{pvj}^{PVJ}) = \beta_p^P + \beta_v^V + \beta_j^J + \beta_{pv}^{PV} + \beta_{pj}^{PJ} + \beta_{vj}^{VJ} + \beta_{pvj}^{PVJ}$$

现在考虑模型

$$\log G_{pvj} = \beta + f(\beta_p^P, \ldots, \beta_{pvj}^{PVJ}) + \lambda_\beta \sum_{pvj} f^{||}(\beta_p^P, \ldots, \beta_{pvj}^{PVJ}) \tag{6}$$

其中 $f^{||}(\beta_p^P, \ldots\ldots, \beta_{pvj}^{PVJ}) = |\beta_p^P| + |\beta_v^V| + |\beta_j^J| + |\beta_{pv}^{PV}| + |\beta_{pj}^{PJ}| + |\beta_{vj}^{VJ}| + |\beta_{pvj}^{PVJ}|$，而 λ_β 为惩罚参数；这里采用对数线性模型的记号，以对任意大小和维度的列联表进行建模。因此，例如， β_{pv}^{PV}表示具有患者特征 p 的患者携带V基因 v 、相较于V基因家族的参照类别的优势比。这是一个饱和的高维L1惩罚对数线性模型；由于施加了惩罚，在仅有一部分VJ相互作用（表明存在富集）显著——无论是否伴有患者表型——这一假设下，可以得到稀疏解。完全饱和的参数化意味着模型包含 $N_P \cdot N_V \cdot N_J$ 个参数，分别对应患者特征类别数、V基因数和J基因数，因此在施加惩罚之前，列联表中的每个单元格都可以无误差地建模。我们通过交叉验证估计 λ_β 。确定用于拟合模型的稀疏参数子集后，我们据此拟合模型，并按常规方法确定参数的显著性。

3 结果

我们对97例随机接受MDT或未接受MDT的前列腺癌患者在基线和随访期间观测到的62662个有功能T细胞受体，以及在基线加两次随访分析中104次观测到的86381个有功能TCR，拟合了纵向贝叶斯混合模型。除非另有说明，所示结果均为基线-随访分析的结果；例如，在动态克隆按MDT分层的比较中，考虑到基线加两次随访分析所需的随机缺失（MAR）假设。T细胞受体–β 互补决定区（CDR）3区域采用Adaptive Technologies的immunoSEQ检测进行测序，其中外周血样本取自患者各次访视时的样本。我们过滤掉无功能序列的T细胞受体，以及在所有随访中的 <8 总读取计数低于阈值的T细胞受体，因此，即使某个克隆在某个横截面中的读取数很少，但经历了显著扩增，仍会被纳入分析。该阈值的设定旨在过滤掉可能缺乏生物学相关性的克隆，同时排除在对低计数进行混合模型建模时产生的数值伪影。哈密顿蒙特卡洛拟合过程收敛后，[3, 27], 我们根据动态成分归属的某一阈值概率，对每位患者中的扩增克隆、收缩克隆和静态克隆进行计数；我们将该阈值设为0.95，以便将动态性更强的扩增或收缩克隆标记为相应类别，因为我们观察到，在定义更严格时，这些克隆表现出更强的关联。

3.1 预后模型

我们对前列腺癌队列的无进展生存期（PFS）终点拟合了Cox比例风险模型；在接受immunoSEQ检测并有基线和随访访视的97名受试者中，观察到53个事件 [8, 21, 25]。由于扩增和收缩数量的计数指标呈右偏分布，并且为了使推断不易受高杠杆观测值影响，我们对扩增数和收缩数进行了对数转换。为提高估计效率并考虑疾病和体能状态的异质性，我们在Cox模型中对患者年龄、雄激素剥夺疗法（ADT）和病灶数量进行了校正。

表2所示的三个Cox模型表明，与动态克隆（扩增和收缩合并）或收缩相比，克隆扩增往往在预后模型中更具显著性；扩增越多，越可能对PFS事件具有保护作用，但其显著性取决于模型是否针对MDT进行了调整；如果模型具有更高的检验效能，则预期纳入MDT的模型会显示出更高的显著性。在生存时间分析中，MDT是解释变量中预后价值最高且保护作用最强的变量。我们还发现，纳入收缩后，扩增往往具有更高的显著性，这表明在收缩的各个分层内，扩增具有更强的保护作用。其解释是，二者的比值对于解释PFS中的变异十分重要，但对这一假设进行正式检验较为困难，且检验效能不足，因此未得到显著结果。最后，MDT混淆了扩增与PFS之间的关系，但对参数估计的考察表明，扩增克隆数量对MDT-PFS关系可能存在轻度的中介作用。

为考察边际关联并采用非参数方法，我们使用Kaplan-Meier估计量估计按扩增数多于或少于30次分层的PFS，所得Log-rank检验的p=0.08（图1）。选择该阈值是为了反映扩增数少于该数值的总体众数，并且其约处于克隆扩增计数的第70百分位数。当在模型中将患者的扩增计数作为连续变量处理时，我们观察到略具更显著性的结果（p=0.06）。

	$\widehat{HR}$	95%置信区间	Z统计量	Pr(>|Z|)
完整模型				
MDT	0.321	(-1.8, -0.5)	-3.51	0.000449
对数扩张	0.825	(-0.46, 0.075)	-1.41	0.158
对数收缩	1.19	(-0.13, 0.48)	1.15	0.251
仅动态性				
对数扩张	0.743	(-0.56, -0.033)	-2.2	0.0275
对数收缩	1.11	(-0.19, 0.4)	0.711	0.477
排除扩张				
MDT	0.299	(-1.8, -0.6)	-3.87	0.00011
对数收缩	1.03	(-0.19, 0.25)	0.268	0.789

表2：包含扩增/收缩与 MDT 组合的三个无进展生存期 Cox 模型，均根据年龄、雄激素剥夺疗法和病灶组进行调整。

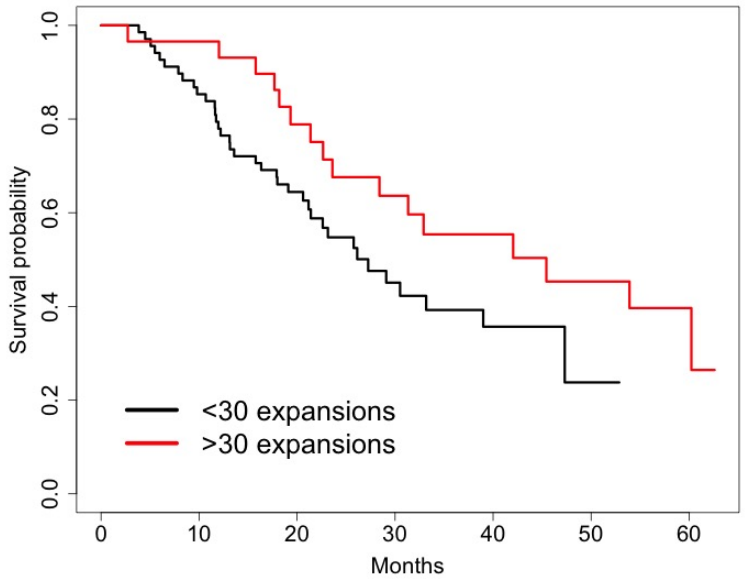


图1：按扩增数量分层的无进展生存期 Kaplan–Meier估计。二分分层的 Log-rank检验 得到 P值 0.08；在 Cox模型 中将其作为连续变量处理时，P值为 0.06。

3.2 VJ基因家族富集与克隆扩增

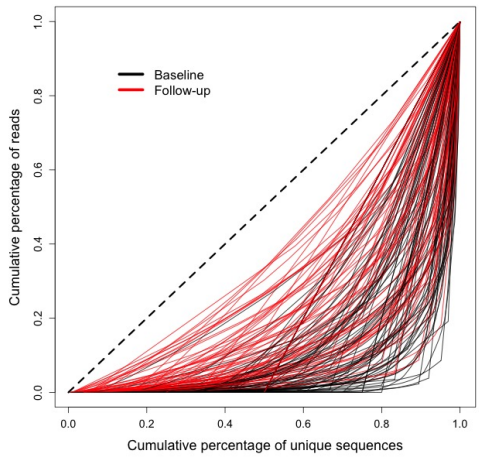
我们将惩罚饱和对数线性模型拟合于V和J基因家族，以及克隆是否被判定为扩增、收缩或动态克隆，结果见表3。由于仅有少数TCR具备其基因家族信息，我们未将D基因纳入分析。模型拟合结果表明，特定V和J基因家族之间存在显著高于预期的富集，而克隆收缩状态不存在富集。表3中扩增状态的主效应被估计为一个较大的负数，这是因为扩增克隆仅占克隆总体的少数。由于所分析的TCR数量达数千个，该分析具有很高的检验效能，因此我们重点关注效应量参数 β所提供的正交信息^；这些参数有助于通过倍数变化理解差异的幅度。大多数^ β为正，这表明当存在基因家族相互作用时，联合频率高于无相互作用零假设下的预期值，而不是低于预期；尤其可以观察到，J基因家族02与V基因家族11和24的组合具有其中一些最高水平。我们确实观察到扩增克隆状态与V基因家族06之间存在显著相互作用；在此情况下，其共现低于根据边际频率所预期的水平，但该现象的生物学依据尚不明确。

	$\hat{\beta}$	95%置信区间	Z统计量	Pr(>|Z|)
扩增主效应				
扩增	-2.95	(-3, -2.9)	-144	0
基因家族相互作用				
TCRBJ02 * TCRBV02	-0.0321	(-0.13, 0.066)	-0.643	0.52
TCRBJ02 * TCRBV03	0.166	(0.055, 0.28)	2.94	0.00328
TCRBJ02 * TCRBV04	0.203	(0.11, 0.3)	4.26	2.02e-05
TCRBJ02 * TCRBV05	0.202	(0.12, 0.28)	4.88	1.05e-06
TCRBJ02 * TCRBV06	0.114	(0.035, 0.19)	2.82	0.00483
TCRBJ02 * TCRBV07	0.263	(0.18, 0.34)	6.57	5.13e-11
TCRBJ02 * TCRBV09	0.327	(0.22, 0.43)	6.04	1.59e-09
TCRBJ02 * TCRBV10	0.163	(0.047, 0.28)	2.76	0.00585
TCRBJ02 * TCRBV11	0.547	(0.44, 0.65)	10	1.34e-23
TCRBJ02 * TCRBV12	-0.0939	(-0.19, 0.0043)	-1.87	0.0609
TCRBJ02 * TCRBV14	0.191	(0.036, 0.35)	2.41	0.0161
TCRBJ02 * TCRBV16	-0.557	(-1, -0.13)	-2.54	0.0112
TCRBJ02 * TCRBV19	-0.227	(-0.32, -0.13)	-4.71	2.45e-06
TCRBJ02 * TCRBV20	0.167	(0.08, 0.25)	3.74	0.000182
TCRBJ02 * TCRBV24	1.57	(1.4, 1.7)	20.9	1.19e-96
TCRBJ02 * TCRBV27	0.242	(0.14, 0.35)	4.51	6.41e-06
TCRBJ02 * TCRBV28	-0.0553	(-0.16, 0.049)	-1.04	0.299
TCRBJ02 * TCRBV29	0.129	(0.00018, 0.26)	1.96	0.0497
TCRBJ02 * TCRBV30	-0.395	(-0.52, -0.27)	-6.07	1.26e-09
基因家族扩增相互作用				
TCRBV06 * 扩增	-0.168	(-0.29, -0.048)	-2.71	0.00678

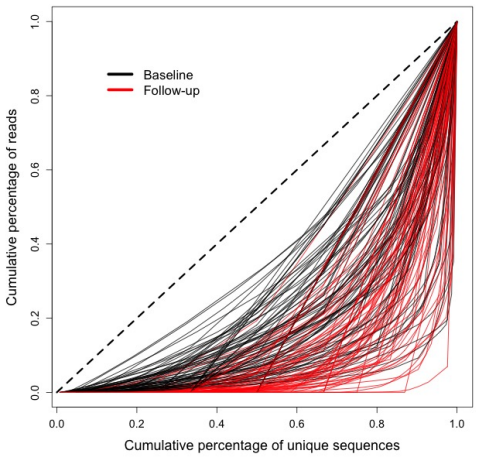
表3：通过对数线性模型惩罚回归识别出的具有扩增相互作用的VJ基因。未显示截距和基因家族主效应。

3.3 克隆扩增和收缩分布

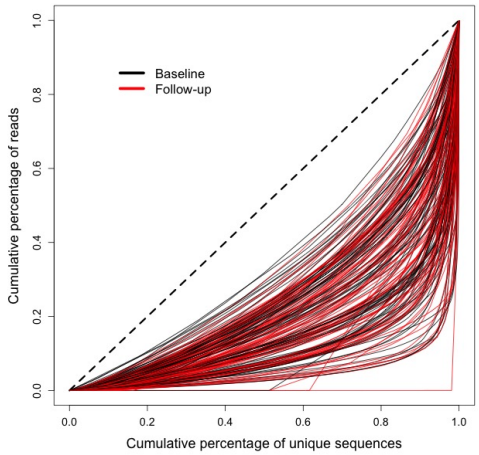
考察 TCR克隆型的纵向变化时，可以分析扩增克隆和收缩克隆的变化是按照某个比例因子进行缩放，还是增加或减少一个绝对量。缩放可能意味着，强烈的免疫反应会维持某些克隆子集原有的相对比例；而按绝对量平移，则在极限情况下会使这些克隆趋向相同的频率。与频率分布熵相关的指标有多种，包括基尼系数。图2中，我们使用并绘制了用于推导这些指标的洛伦兹曲线，并按患者和纵向随访时间分层，分别展示扩增克隆、收缩克隆和静态克隆的情况。曲线按基线-随访阶段呈现出明显分层，且扩增克隆的频率分布趋于更加均匀，这意味着克隆的变化平均而言更可能是平移而非缩放——起始频率较高或较低的克隆，改变的是 X 这一绝对量，而非 X%。



(a) 扩增克隆洛伦兹曲线。



(b) 收缩克隆洛伦兹曲线。



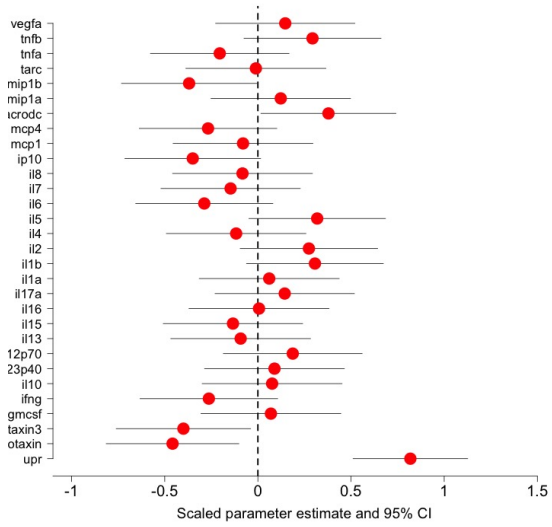
(c) 静态克隆洛伦兹曲线。

图2：按基线/随访横截面分层的受试者中，扩增、收缩和静态克隆的洛伦兹曲线

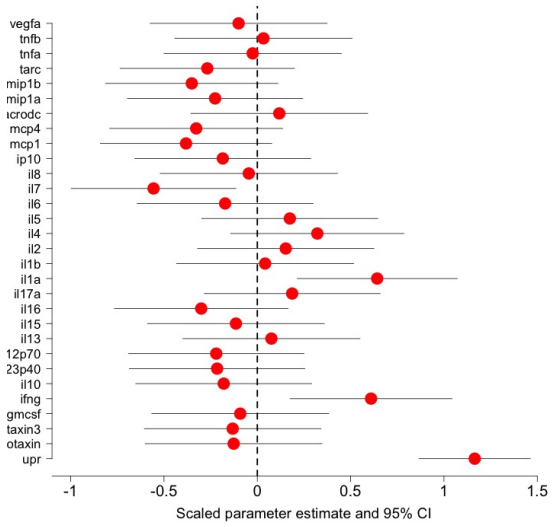
3.4 基线生物标志物与克隆动态性

我们拟合线性回归模型，分析扩增和收缩的数量与基线生物标志物之间的关联，并在MDT患者和非MDT患者的分层内分别进行分析。图3和图4展示了标准化边际关联参数估计。克隆动态变化与生物标志物之间最主要且明确的关联是UPR；这一关联不因扩增或收缩，也不因MDT或非MDT状态而显著变化。我们还发现了其他较弱的关联（例如 Il7），但未通过多重检验校正。此外，标准化效应量在扩增和MDT分层间也存在一定程度的变异。

我们利用 glinternet（一种用于在惩罚型高维回归情境下识别这些项、并保持主效应模型层级排序的模型和工具）探索了基线生物标志物之间的相互作用 [18]。结果见表4，显示在受试者 [28] 中，gmcsf * mip1b 和 il15 * vegfa 与扩增数量之间存在两种较弱的相互作用。最后，在图5中，我们发现 MDT 分层在基线-随访分析中表现出显著更多的收缩和扩增；而在基线以及两次随访分析中，动态克隆总体上也更多。我们还发现，在有 HLA 分型数据的受试者子集中，HLA等位基因 A*03 和 C*07 的存在分别与克隆扩增呈正相关和负相关，结果见 A3节。

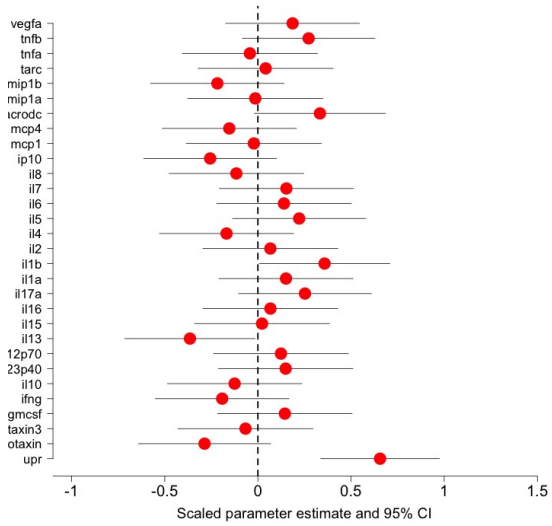


（a）MDT治疗观测值中，基线生物标志物与扩增之间的边际关联。

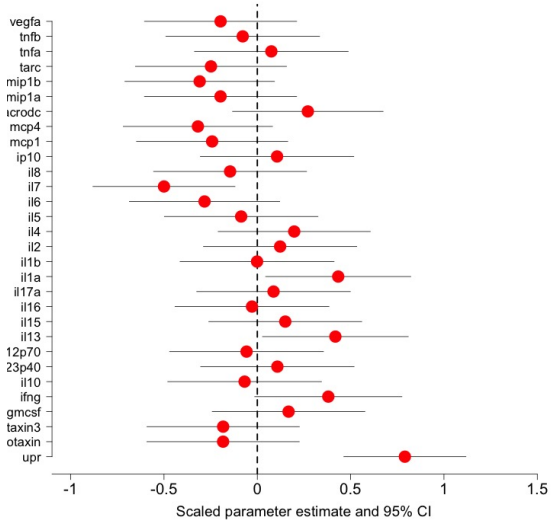


（b）非MDT观测值中，基线生物标志物与扩增之间的边际关联。

图3：按MDT分层的基线生物标志物与克隆扩增之间的边际关联。



（a）MDT治疗观测值中，基线生物标志物与收缩之间的边际关联。

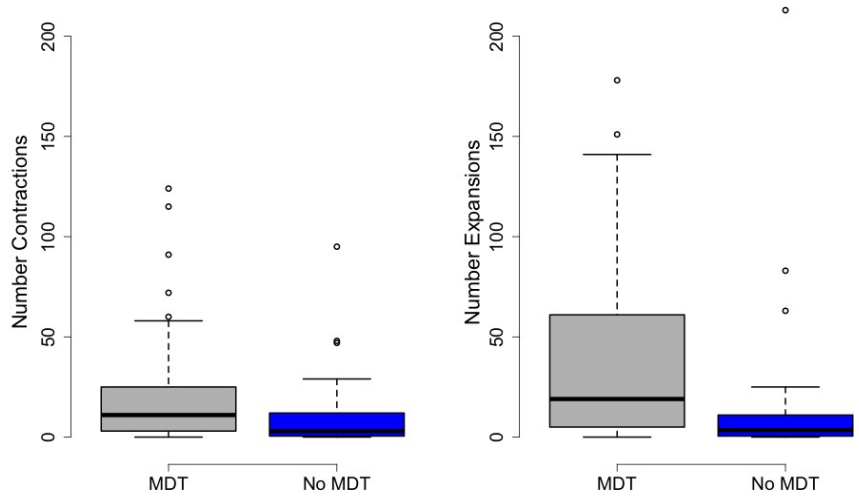


（b）非MDT观测值中，基线生物标志物与收缩之间的边际关联。

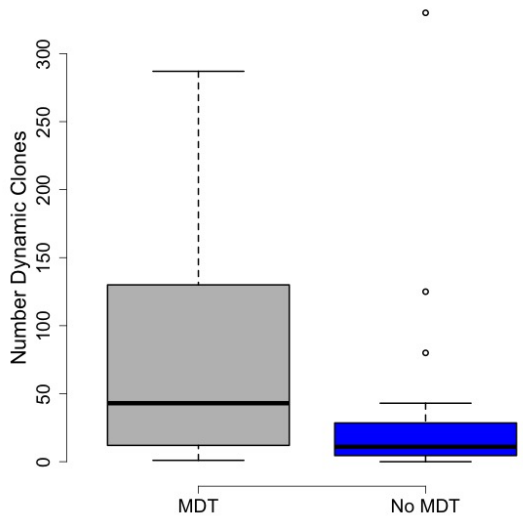
图4：　基线生物标志物与克隆收缩之间的边际关联，按MDT 分层。

		$\hat{\beta}$	95%置信区间	t统计量	Pr(>|t|)
主效应					
	upr	2.93e-05	(2.2e-05, 3.6e-05)	8.28	1.42e-12
	粒细胞-巨噬细胞集落刺激因子	7.8	(-0.62, 16)	1.84	0.0689
	il10	0.869	(0.092, 1.6)	2.22	0.0289
	il1a	0.0236	(-0.017, 0.064)	1.17	0.247
	il15	-0.516	(-0.94, -0.09)	-2.41	0.0181
	vegfa	-0.0351	(-0.067, -0.0033)	-2.19	0.0311
	mip1b	0.00516	(-0.0054, 0.016)	0.971	0.334
相互作用					
	粒细胞-巨噬细胞集落刺激因子 * mip1b	-0.1	(-0.2, -0.0053)	-2.1	0.0388
	il15 * vegfa	0.0119	(0.002, 0.022)	2.39	0.0192

表4: 采用交互作用导向的惩罚回归，将克隆扩增对基线生物标志物进行回归 d回归 在强制执行层级排序模型时



(a) 按 MDT 分层的第 2 次随访分析中收缩（左侧）和扩增（右侧）的箱线图。



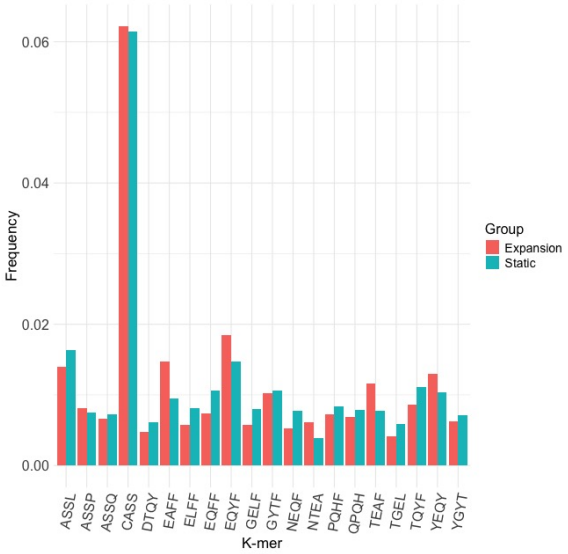
(b) 按 MDT 分层的第 3 次随访分析中动态克隆的箱线图。

图5: 第 2 次和第 3 次随访分析中的扩增、收缩和动态克隆，分别如上所述 . 所有 p 值在广义线性模型中，针对 MDT 分层 $< 10^{-8}$ 。

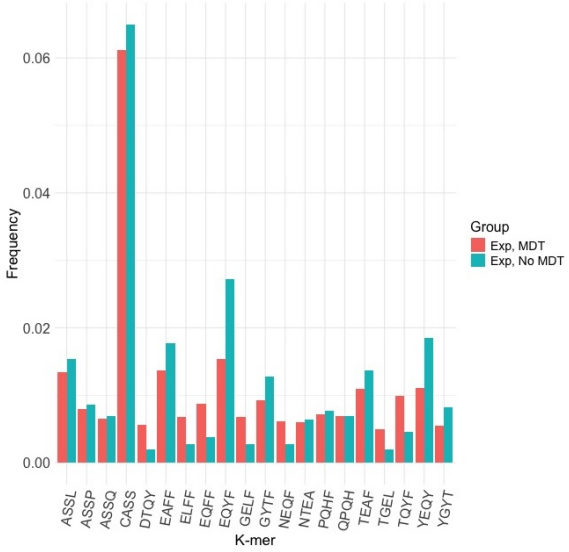
3.5 基序和核苷酸分析

基序分析的可视化结果显示，扩增克隆与静态克隆中常见基序的相对比例相似（图6a），这表明扩增克隆中并未富集某一种或另一种基序。相比之下，对按是否接受MDT治疗分层的扩增克隆中的基序相对比例进行目视检查后发现存在差异：MDT诱导的扩增中，各基序的分布更均匀。这一结果可能支持这样一种观点：MDT诱导的扩增克隆具有某些不同于未接受MDT时出现的基序特征。分布差异的正式检验未发现显著差异，但基序类别数量众多，导致分析检验效能不足。尽管

检验效能充足，我们发现，就克隆型聚类而言，扩增、收缩和静态克隆之间的CDR3长度没有差异。同样，在按个体匹配所比较的克隆数量后，通过t检验和Wilcoxon秩和检验发现，扩增克隆与静态克隆的平均核苷酸编辑距离仅存在轻微差异（p=0.25）；扩增克隆的平均编辑距离略大。



(a) 动态克隆中最常见的基序，按其属于扩增克隆或静态克隆进行分层。



(b) 扩增克隆中最常见的基序，按其是否为MDT患者与非MDT患者中的扩增克隆进行分层。

图6：按扩增状态和MDT分层的常见基序。

4 讨论

在本研究中，我们探讨了MDT如何影响TCR库及其纵向变化，以及这些变化如何影响患者反应——这种反应通过无进展生存期进行衡量，无论是否结合MDT。此外，我们还尝试刻画纵向克隆轨迹的性质，无论其表现为平移还是缩放，并探讨将其判定为克隆扩增或收缩是否可能与特定基因家族、基线生物标志物和受体基序的富集相关。

我们分析得出的一个重要且具有验证性的见解是，克隆扩增可能与患者预后存在一定关联；这种关联在边际分析中存在，而在MDT分层内则较弱，同时还需考虑合并症指标并进行校正。这提示，就扩增克隆型的数量而言，具有更强或更广泛免疫反应的患者可能会较晚出现疾病进展。由于本研究较为明确的信号之一是，随机MDT显著增加了扩增克隆的数量，因此，对其与无进展生存期之间 modest 关联的解读应谨慎，因为该关联受到MDT的混杂影响。尽管我们的研究构成了一个规模相对较大的TCR测序个体前瞻性队列，但对于97例患者中发生的53个无进展生存期事件而言，我们的检验效能仍然不足，因此有必要在未来规模更大的队列中开展分析。

许多重要的方法学研究聚焦于VDJ基因，而我们的方案则朝着系统评估T细胞受体VJ基因家族之间的富集迈出了重要一步。由于我们的框架具有普适性，还可以进一步分析克隆型的其他特征与基因家族之间的相互作用，例如该克隆型是扩增、收缩还是保持稳定。我们的分析揭示了若干V基因家族与J基因家族之间有趣的富集关系，以及扩增克隆与某个特定V基因家族之间的相互作用。后一项结果结合图6b所示的趋势，表明对MDT有反应时扩增的克隆型与未接受MDT时扩增的克隆型之间存在细微差异。

5 补充材料

补充材料包含模拟结果（见A1节），证实参数估计与生成数据时所依据的参数一致；还包括纵向模型混合成分的推导（见A2 节），以及HLA分型分析（见A3节）。

参考文献

[1] Alan Agresti。 *Categorical data analysis*。John Wiley & Sons，2013 年。

[2] Meriem Bensouda Koraichi、Silvia Ferri、Aleksandra M. Walczak 和 Thierry Mora。推断健康个体的 T 细胞受体库动态。 *Proceedings of the National Academy of Sciences*，120 4 2207516120，2023 0027 8424，1091 6490 101073 2207516120（）： e January。ISSN - -。doi:。/pnas.。URL https://pnas.org/doi/10.1073/pnas.2207516120。

[3] Michael Betancourt 和 Mark Girolami。用于层次模型的哈密顿蒙特卡洛。 *Current trends in Bayesian methodology with applications* 79 30 2 4, 2015 ，（）：–。出版商：CRC Press Boca Raton，FL。

[4] George Casella 和 Roger Berger。 *Statistical inference*。CRC press，2024。

[5]Lin Chen、Yuan Hu、Bohao Zheng、Limei Luo 和 Zhenzhen Su。癌症中的人类 TCR 库。*Cancer Medicine*，13(17): e70164，2024年9月。ISSN 2045-7634，2045-7634。doi: 10.1002/cam4.70164。URL https://onlinelibrary.wiley.com/doi/10.1002/cam4.70164。[6] Yi-Tung Chen、Hung-Chih Hsu、Yun-Shien Lee、Hsuan Liu、Bertrand Chin-Ming Tan、Chia-Yin Chin、Ian Yi-Feng Chang 和 Chia-Yu Yang。在化疗期间对转移性结直肠癌患者进行T细胞受体库纵向高通量测序，揭示外周血TCR多样性的动态变化及其预后意义。 *Frontiers in Immunology* 12 743448, 2022 1664 3224 10 3389 2021 743448 ，：1月。ISSN -。doi:。/fimmu。。URL https://www.frontiersin.org/articles/10.3389/fimmu.2021.743448/full。[7] Yongsheng Chen、Wenrui Xiao、Shiyi Yuan、Cong Wang、Meizhen Shi、Dan Yu、Ying Zhang 和 Shifeng Lou。继发性成人HLH淋巴细胞中TCR免疫受体库的特征及特异性差异基因分析。 *Immunology Letters*，275:107023，2025。[8] David R Cox。回归模型与生命表。 *Journal of the Royal Statistical Society: Series B (Methodological)* 34 2 187 202, 1972 ，(): –。

[9] William S DeWitt、Ryan O Emerson、Paul Lindau、Marissa Vignali、Thomas M Snyder、Cindy Des-marais、Catherine Sanders、Heidi Utsugi、Edus H Warren、Juliana McElrath 等。针对急性病毒感染模型的细胞毒性T细胞反应动态。 *Journal of virology* 89 8 4517 4526, ，(): –。2015 出版商：Am Soc Microbiol。[10] Ronald A Fisher。关于列联表中x2的解释及P值的计算。 *Journal of the royal statistical society*，85(1):87–94，1922。出版商：JSTOR。[11] Ronald Aylmer Fisher。 *Statistical methods for research workers*。第5卷。Oliver and Boyd，1928。[12] Mark Girolami 和 Ben Calderhead。黎曼流形朗之万与哈密顿蒙特卡洛方法。 *Journal of the Royal Statistical Society Series B: Statistical Methodology*，73(2):123–214，2011。出版商：Oxford University Press。[13] Liping Guo、Xingang Bi、Yajian Li、Li Wen、Wen Zhang、Weixing Jiang、JianHui Ma、Lin Feng、Kaitai Zhang 和 Jianzhong Shou。肾细胞癌患者TCR受体库分析的特征、动态变化及预后意义。*ogy* 251 1 26 37, 2020 0022 3417，1096 9896 10 1002 5396 *The Journal of Pathol* ，(): – 5月。ISSN - -。doi:。/path。。URL https://pathsocjournals.onlinelibrary.wiley.com/doi/10.1002/path.5396。[14] Matthew D Hoffman、Andrew Gelman 等。无调转采样器：在哈密顿蒙特卡洛中自适应设定路径长度。 *J. Mach. Learn. Res.*，15(1):1593–1623，2014。

[15] Emmi Jokinen、Jani Huuhtanen、Satu Mustjoki、Markus Heinonen 和 Harri L¨ahdesm¨aki。使用TCRGP预测T细胞受体与表位之间的识别。*Biology* 17 3 1008814, 2021 1553 7358 10 1371 1008814 , ( ):e 3月。 *PLOS Computational*

ISSN -。doi:。/journal.pcbi。。URL https://dx.plos.org/10.1371/journal.pcbi.1008814。

[16] Joel Kidman、Nicola Principe、Mark Watson、Timo Lassmann、Robert A. Holt、Anna K. Nowak、Willem Joost Lesterhuis、Richard A. Lake 和 Jonathan Chee。《与免疫检查点疗法成功疗效相关的 TCR 库特征》。*munology* 11 587014, 2020 1664 3224 10 3389 2020 587014 , : 10月。 *Frontiers in Im-*

ISSN -。doi:。/fimmu.。。网址：https://www.frontiersin.org/article/10.3389/fimmu.2020.587014/full。

[17] Nan M. Laird 和 James H. Ware。*Biometrics*，第963–974页，1982年。双向数据的随机效应模型 .

[18] Michael Lim 和 Trevor Hastie . 通过层次化组套索正则化学习相互作用 .
Journal of Computational and Graphical Statistics, 24(3):627–654, 2015年。

[19] Roderick J Little . 缺失数据分析 . *Annual Review of Clinical Psychology*, 20, 2024 .
出版商：Annual Reviews。

[20] Roderick J Little 和 Donald B Rubin . 2019年。 John Wiley & Sons. *Statistical analysis with missing data.*

[21] Ethan B Ludmir、Alexander D Sherry、Bryan M Fellman、Suyu Liu、Tharakeswara Bathala、Cara Haymaker、Marina N Medina-Rosales、Alexandre Reuben、Emma B Holliday、Grace L Smith 等。为寡转移性胰腺导管腺癌患者的全身治疗增加转移灶定向治疗（EXTEND）：一项多中心随机II期试验。 *Journal of Clinical Oncology* 24, 2024 ，页码 JCO–。出版商：Wolters Kluwer Health。

[22] Long Ma、Liwen Yang、Bin Shi、Xiaoyan He、Aihua Peng、Yuehong Li、Teng Zhang、Suhong Sun、Rui Ma 和 Xinsheng Yao。利用高通量测序分析外周T细胞中TCRβ链的V-D-J和V-J重排相关的CDR3库。 *Scientific Reports*, 6(1):29544, 2016年7月。ISSN 2045-2322。doi: 10.1038/srep29544。URL https://www.nature.com/articles/srep29544。

[23] Bernard A Rosner 等。《 *Fundamentals of biostatistics*》，第6卷。Thomson-Brooks/Cole，贝尔蒙特（加利福尼亚州），2006年。

[24] Julie Rytlewski、Shibing Deng、Tao Xie、Craig Davis、Harlan Robins、Erik Yusko 和 Jadwiga Bi- enkowska。用于提高外周血中临床相关扩增T细胞识别特异性的模型。《 *PLOS ONE*》，14(3):e0213684，2019年3月。ISSN 1932-6203。doi: 10.1371/jour- nal.pone.0213684。URL https://dx.plos.org/10.1371/journal.pone.0213684。

[25] Alexander D. Sherry、Tharakeswara K. Bathala、Suyu Liu、Bryan M. Fellman、Stephen G. Chun、Nikesh Jasani、B. Ashleigh Guadagnolo、Anuja Jhingran、Jay P. Reddy、Paul G. Corn、Amishi Y. Shah、Kelsey W. Kaiser、Amol J. Ghia、Daniel R. Gomez 和 Chad Tang。寡转移实体瘤的确定性局部巩固治疗：随机篮式试验 EXTEND 导入阶段的结果。《 *International Journal of Radiation Oncology*Biology*Physics*》，114 5 910 918, 2022 03603016 10 1016 2022 05 023 ( ): – December . ISSN . doi: . /j.ijrobp. . . . URL https://linkinghub.elsevier.com/retrieve/pii/S0360301622004758。[26] Mikhail Shugay、Dmitriy V Bagaev、Ivan V Zvyagin、Renske M Vroomans、Jeremy Chase Crawford、Garry Dolton、Ekaterina A Komech、Anastasiya L Sycheva、Anna E Koneva、Evgeniy S Egorov、Alexey V Eliseev、Ewald Van Dyk、Pradyot Dash、Meriem Attaf、Cristina Rius、Kristin Ladell、James E McLaren、Katherine K Matthews、E Bridie Clemens、Daniel C Douek、Fabio Luciani、Debbie van Baarle、Katherine Kedzierska、Can Kesmir、Paul G Thomas、David A Price、Andrew K Sewell 和 Dmitriy M Chudakov。VDJdb：一个收录已知抗原特异性T细胞受体序列的精选数据库。《 *Nucleic Acids Research*》，46(D1):D419–D427，2018年1月。ISSN 0305-1048, 1362-4962。doi: 10.1093/nar/gkx760。URL http://academic.oup.com/nar/article/46/D1/D419/4101254。[27] STAN 开发团队。《Stan 参考手册》，2024年。URL https://mc-stan.org。[28] Robert Tibshirani。通过套索回归实现回归收缩与变量选择。《 *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58（1）:267 88, 1996.

[29] Sara Valpione、Elena Galvani、Joshua Tweedy、Piyushkumar A. Mundra、Antonia Banyard、Philippa Middlehurst、Jeff Barry、Sarah Mills、Zena Salih、John Weightman、Avinash Gupta、Gabriela Gremel、Franziska Baenke、Nathalie Dhomen、Paul C. Lorigan 和 Richard Marais。单个疗程免疫治疗后外周 T 细胞动态变化揭示免疫唤醒。*Nature Cancer* 1 2 210 221, 2020 2662 1347 10 1038 43018 019 0022 , ( ): – February 。

ISSN - 。doi: 。/s - - -x。URL https://www.nature.com/articles/s43018-019-0022-x。

[30] Qian Wang、Delong Feng、Yang Song、Zhi Hu、Qianjin Lu 和 Ming Zhao。《系统性自身免疫性疾病中 IgG 库的 V/D/J 基因偏倚使用与克隆多样性和疾病活动度及临床特征的相关性》。《 *Immunological Investigations*》，第 1–21 页，2025 年。

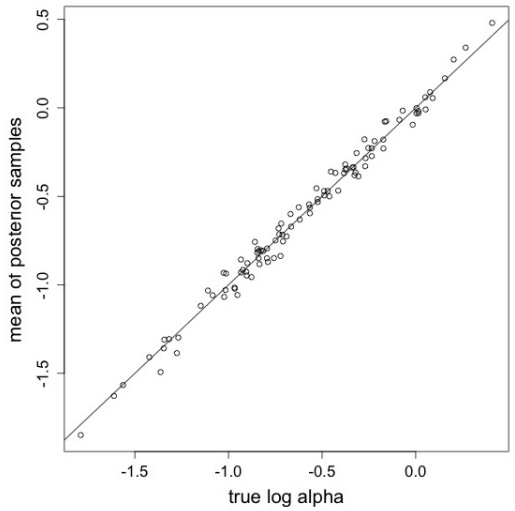
补充材料：利用纵向贝叶斯混合模型识别扩增的 TCR克隆型及其与癌症患者预后、转移灶定向治疗和 VJ基因富集的关联David Swanson

，Alexander Sherry ，Cara Haymaker ，Alexandre Reuben ，Chad Tang

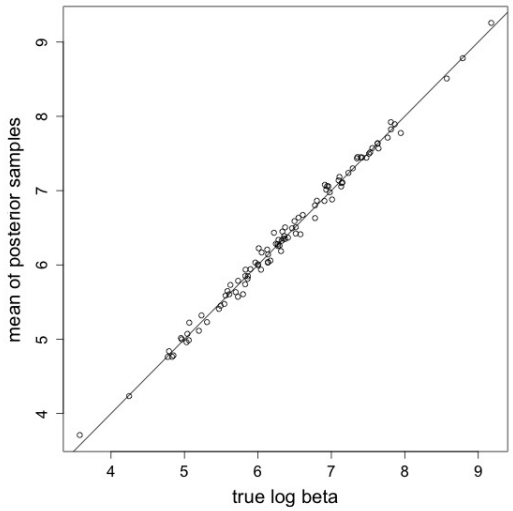
A1 模拟

我们进行了多项模拟，以考察拟合过程能否恢复生成样本所依据的参数。我们将 α_i 和 β_i 的 100 个对数值样本视为服从均值为 -0.5 或 -0.3 、标准差分别为 6.0 或 4.4 的高斯分布。在这些参数化下，标准差分别为 0.5 或 0.3。随后进行指数变换后，这些参数对应的伽马先验均值大致相同，但 -0.3 和 4.4 的组合具有更高的平均方差，泊松均值参数便从该方差中抽取。直观而言，这应当能提供更多纵向信息，以判定克隆属于动态成分还是静态成分。根据模拟设置，我们生成了 60,000 个克隆，每个克隆在 2 或 3 个随访时间点进行观测，其中 20% 在动态模型下生成，80% 在静态模型下生成。在 3 个随访时间点的模拟中，约 34% 的样本在第 2 或第 3 个随访时间点存在缺失。偏移项服从形状参数为 5000/20、率参数为 1/20 的伽马分布。

模型采用 2 次或 3 次随访的组合进行拟合；在后一种情况下，存在缺失，并使用方差较小或较大的伽马分布的 α_j 和 β_j 参数化。我们通过 STAN 使用哈密顿蒙特卡洛（HMC）拟合模型，在 2 条独立链上各抽取 2500 个样本，预热期为 500 [27]。哈密顿蒙特卡洛的一个优点是样本几乎相互独立，因此能更快地将抽样误差降至比估计误差小一个数量级。

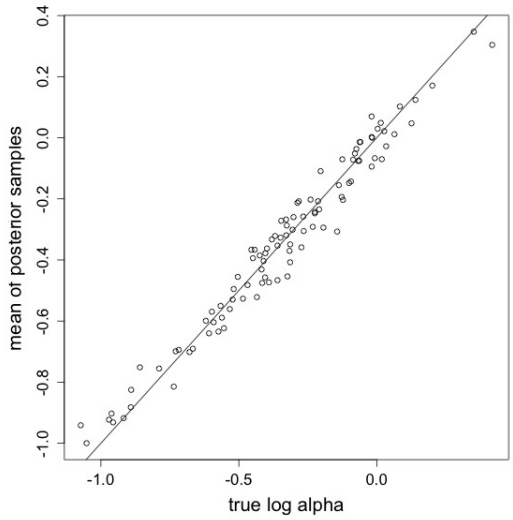


(a) α 在 3 次随访中的小方差模拟。

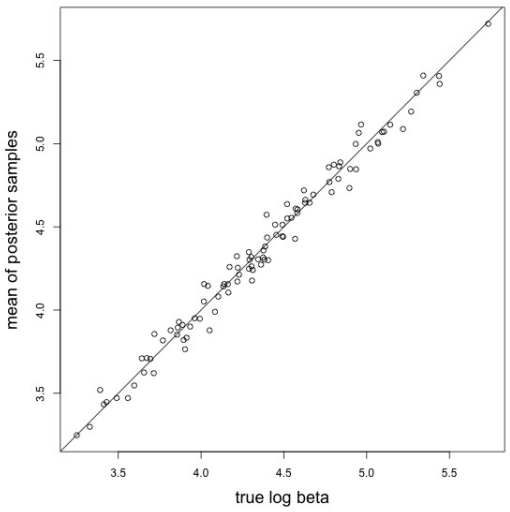


(b) β 在 3 次随访中的小方差模拟。

图 A1：真实参数与后验样本均值的对比，3 次随访。45 度线表示完全一致。

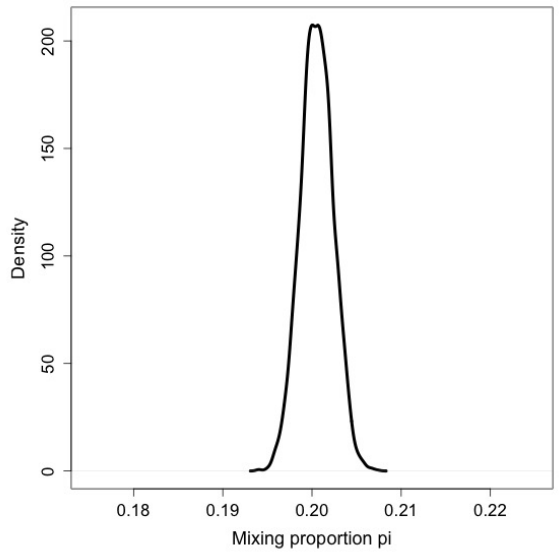


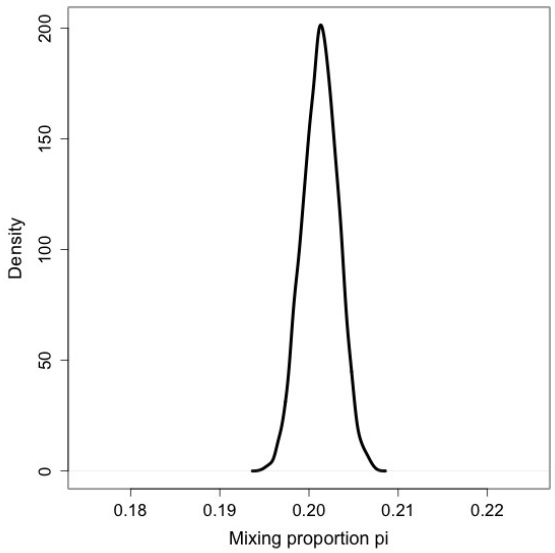
(a) 在 2 次随访中对 α 进行的大方差模拟。



(b) 在 2 次随访中对 β 进行的大方差模拟。

图 A2：真实参数与后验样本均值的对比，共 2 次随访。45 度线表示完全一致。



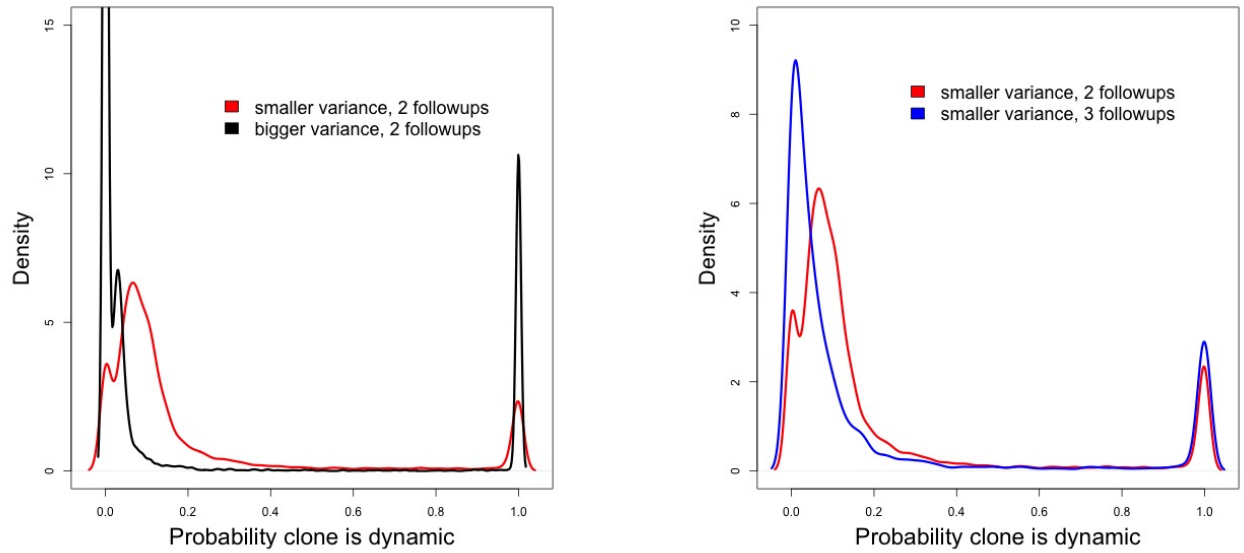


(a) 在较大方差参数化且进行 2 次随访的模拟中， π 样本的分布。(b) 在较小方差参数化且进行 3 次随访的模拟中， π 样本的分布。

图 A3：不同模拟情景下的后验 π 样本。

模拟结果表明，拟合过程能够很好地刻画数据生成机制中使用的参数；图 A1 和图 A2 显示，真实的 α_i 和 β_i 与后验样本均值高度吻合。同样，图 A3 表明，logit π 的样本经逆 logit 变换后与真实参数 0.2 高度吻合，其分布几乎完美地以.....为中心

值。轻微的偏差可能是由于在应用反函数之前，先在对称 logit 空间中拟合参数所致。



(a) 在方差较大和较小的情况下，随访 2 次时克隆的成分归属。(b) 在方差较小的情况下，随访 2 次和 3 次时克隆的成分归属。

图 A4：不同模拟情景下平均概率成分归属的密度图。

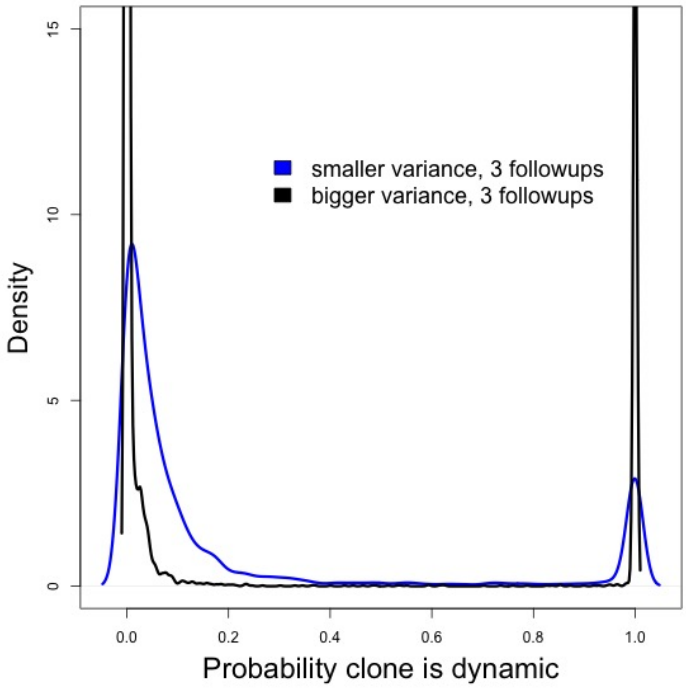


图 A5：3 次随访下、较小和较大方差时克隆成分归属的密度。

图 A4 和图 A5 同样显示，该模型能够按预期良好地区分静态克隆和动态克隆。每个密度图中左侧的峰对应静态克隆，而右侧的峰对应动态克隆。各图中峰的尖锐程度也如预期般有所不同——与 2 次随访相比，3 次随访提供了更多信息来区分动态行为和静态行为，因此峰的尖锐程度更高。同样，在使用 Gamma先验、且后续泊松均值参数平均具有更大变异的模拟情景中，由于动态行为更容易识别，在其他条件相同的情况下，峰的尖锐程度也更高。

A2 静态和动态混合成分的推导

A2.1 推导静态模型

由于对于静态成分，我们有

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

and

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

那么，似然的概率质量函数（pmf）和先验的密度为

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

$$p(\lambda_{ij} | D_{ij} = 0, \alpha_j, \beta_j) = \frac{\beta_j^{\alpha_j}}{\Gamma(\alpha_j)} \lambda_{ij}^{\alpha_j - 1} \exp\left(-\beta_j \lambda_{ij}\right),$$

分别为。将这些项相乘并对 λ_{ij} 积分消去后得到

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

其中，我们利用共轭性计算积分，并识别出以更新后的参数 $\alpha_j' = \sum_k C_{ijk} + \alpha_j$ 和 $\beta_j' = \sum_k O_{jk} + \beta_j$为参数的 Gamma分布核。关于 λ_{ij} 的积分给出了 Gamma(α_j', β_j') 的归一化常数的倒数。

A2.2 推导动态模型

对于动态成分的克隆似然，我们有

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

对于先验，则有

$$\lambda_{ijk}|(D_{ij}=1) \sim \mathrm{Gamma}(\alpha_j, \beta_j)$$

因此，对于每个 k，概率质量函数和密度表达式分别为

$$l(C_{ijk}|\lambda_{ijk},\, O_{j\cdot}) = \frac{\exp(-\lambda_{ijk}\, O_{jk}) \lambda_{ijk}^{C_{ijk}} O_{jk}^{C_{ijk}}}{C_{ijk}!}$$

$$p(\lambda_{ijk}|D_{ij}=1, \alpha_j, \beta_j) = \frac{\beta_j^{\alpha_j}}{\Gamma(\alpha_j)} \lambda_{ijk}^{\alpha_j-1} \exp\left(-\beta_j \lambda_{ijk}\right),$$

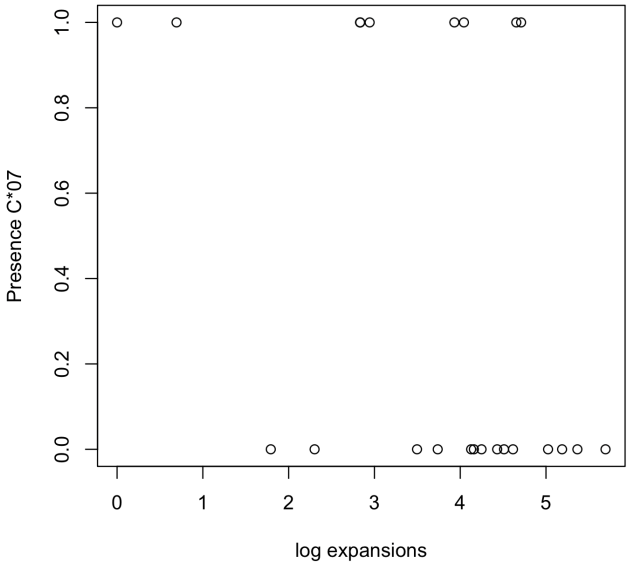
分别为。随后对 k 求积得到

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

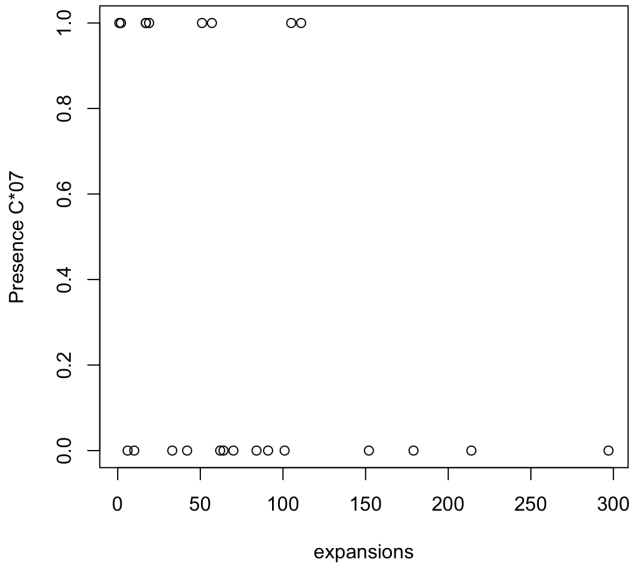
这里再次利用共轭性计算积分，并识别出更新后的参数 $\alpha_j' = \sum_k C_{ijk} + \alpha_j$ 和 $\beta_j' = \sum_k O_{jk} + \beta_j$所对应的 Gamma分布核。对 λ_{ijk} 积分得到 Gamma(α_j', β_j') 的归一化常数的倒数。

A3 HLA分型与克隆扩增

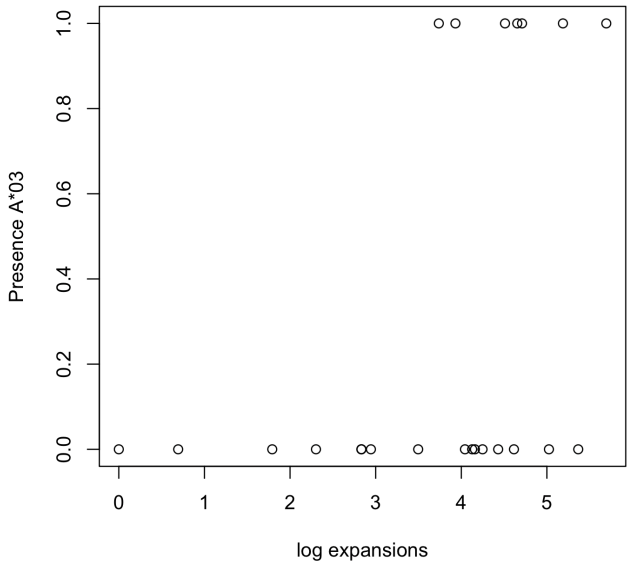
有 23 位患者具备可供分析的 HLA分型和基线-随访数据。结果显示，A*03 和 C*07 等位基因的存在与克隆扩增之间存在较弱的关联。



(a) 对数变换下线性模型的 P值，0.043



(b) 未变换的关联的可视化示意图。



(c) 对数变换下线性模型的 P值，0.040