TLB 的VDJ重新分析（初步）

之前想着是对富集的TLB的BCR序列做抗原预测，但是目前的好像没有合适的工具能完成这个任务。所以现在的目标降低一点，就是分析TLB中VDJ基因的使用情况以及抗体CDR3区域有没有特异之处，这肯定需要健康对照。但现在协和的对照样本还没有到位，并且他们计划只测序一个病人貌似也不是太够，所以我还计划下载10x官网提供的健康人的PBMC的样本做完对照，虽然meta不全，并且可能存在年龄和种族等不可控因素，但是总是比没有强。。

0 vdj测序原理学习

补充了10x vdj测序的相关知识：参考1，参考2

- [x] 为什么要5‘ 端测序：VDJ区域在mRNA的5‘端
- [x] 为什么效率比3’ gene expression 低：文库分成3份分别去测TCR/BCR/RNA
- [x] 如何构建全长的VDJ区域（约650bp，长于单个reads数目）：巢式PCR加上reads mapping

1 工具选择

在免疫组库这个方面，没有一个特别公认的工具，大家一般也就八仙过海，对自己感兴趣的部分简单弄一下就完事了。还算比较系统的工具有

1. scRepertoire：和seurat的整合比较好，用于探索整体的分布比较合适
2. immunarch：从bulk时代就比较权威的免疫组库探索工具，应该比较全面
3. ~~clonotypr：和seurat的整合比较好，用于探索序列比较合适~~

clonotypr年久失修无法安装，先用这剩下两个工具，然后看看其他测过单细胞免疫组的文章的特色分析（类似泽民老师的**startrack**）

2 总体思路

现在已知plasma分泌的自身抗体是SLE疾病进程中的一个重要因素，但是还没有人从单细胞免疫的角度切入（也没有类似的数据）。题外话，由于COV19的压力，单细胞免疫组的应用得到了广泛的推展

从当前我们的scRNA数据看，IgA plasma占据主要（但是缺少对照样本，不能下定论）

1. Plasma中是否能发现非常富集的克隆型（CDR3）[如何定义克隆型？是CDR3区域完全一样还是必须全体的VDJ都一样？需要考虑突变吗？完全一样还是允许1/更多碱基的差异？]
2. IgA 是否相关
3. TLB_mem1中富集的克隆型
4. TLB_plasma是否真实存在，是否有BCR（在当前的数据中没有）

总结一下：谁产生了可能与SLE发病相关的抗体（IgA plasma/IgG plasma/TLB_mem1/TLB_mem2），然后映射到seurat的umap上

剩下的就是免疫组的常见套路：

1. 疾病与健康对照之间的VDJ基因使用偏好情况
2.

3 分析思路

从当前的工具的支持情况看，主要分为两个level

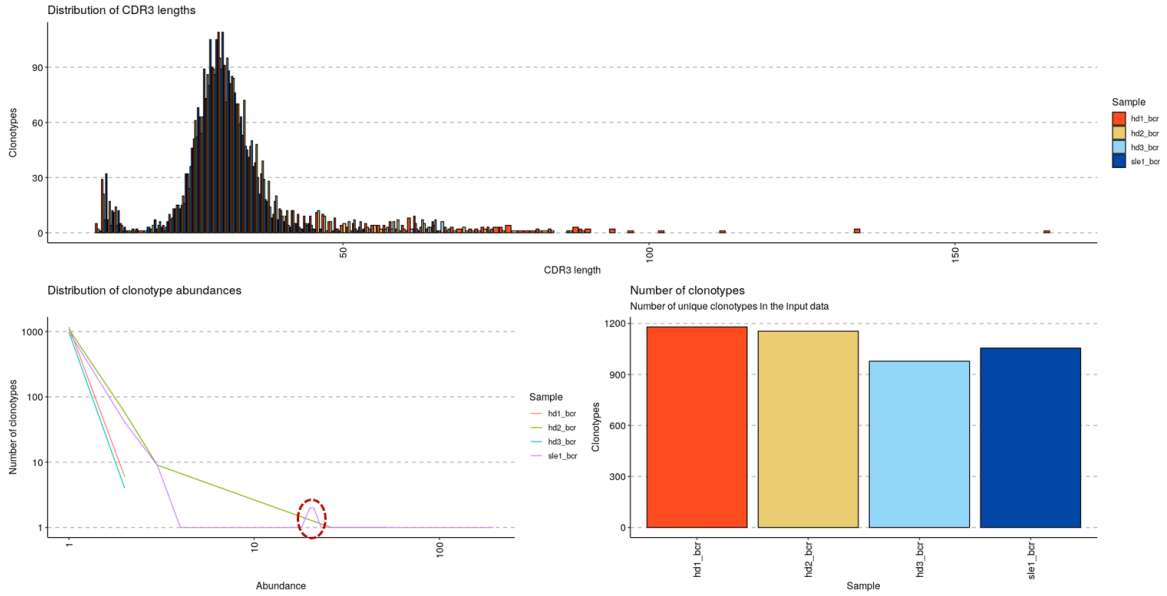
1. 样本level：除了样本一些基本信息的统计，还可以分析样本之间组库多样性/重叠程度的比较，以及特定vdj基因的使用偏好性
2. 克隆型level：有点类似转录组的cluster，可以跨越样本/时间尺度追踪特定克隆型，以及immunarch特有的基于免疫数据库注释和scRepertoire整合的startrack算法

4 初步结果

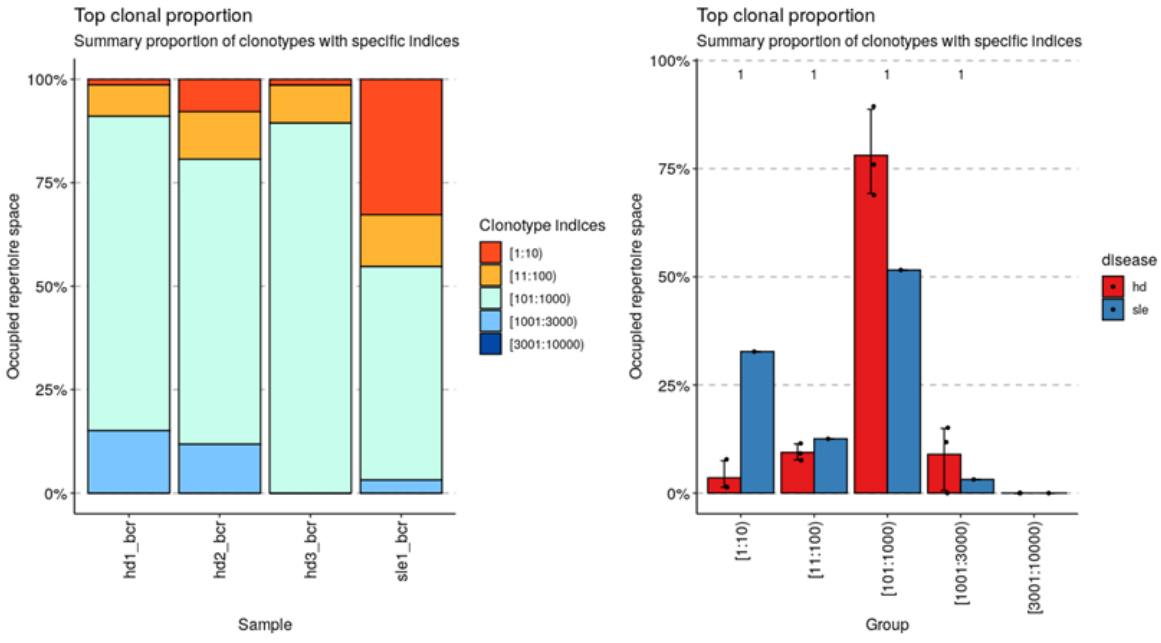
首先是sle数据和10x一个健康样本BCR组库分析，但是发现数目太少对照效果不好。其实我还想找cov19作为非正常对照，但是怎么都找不到。所以又只好从10x官网找来两个健康作为正常对照

4.1 BCR

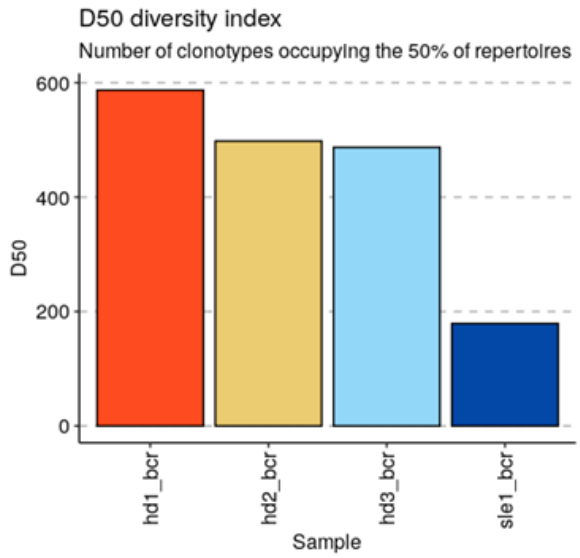
首先我们观察到4个样本测到的BCR克隆数目是差不多的，代表测序数据可比。但左下图可以发现SLE出现了一个明显的BCR克隆型富集

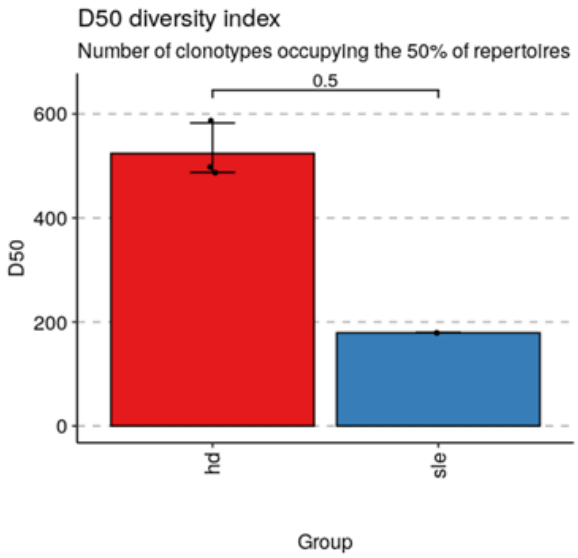


所以特别关注了这些富集的克隆型，可以发现TOP10富集的克隆型在SLE中所占比例要明显高于HD对照（左图是单个样本，右图是分为两组）

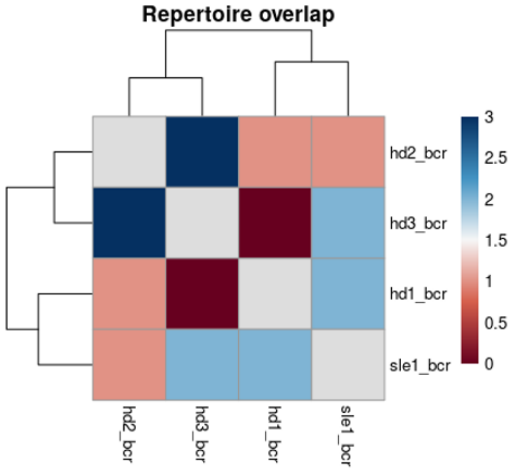


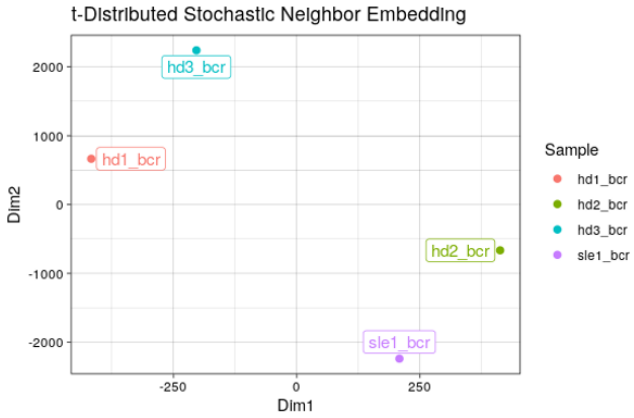
用***D50 多样性指数***证实这一点，发现SLE患者的BCR组库的多样性有显著的下降





接下来按照克隆型对样本聚类（heatmap，tsne），但由于样本太少，看不出规律





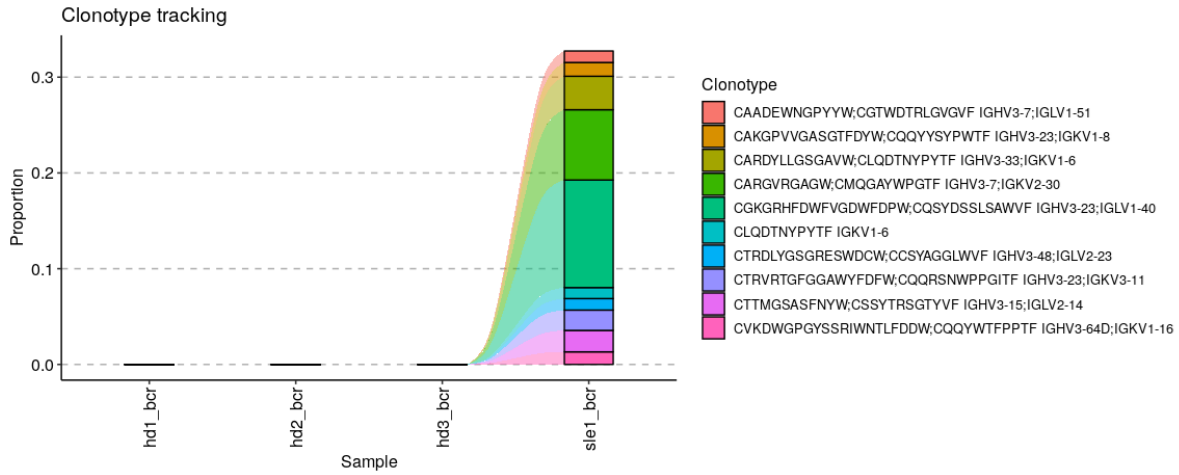
疾病样本之间的公共克隆型可能对疾病发展有着比较重要的作用，但是目前只有一个SLE样本，mark一下，以后补上分析

```
pr.nt <- pubRep(sle_bcr$data, "nt", .verbose = F)
pr.aav <- pubRep(sle_bcr$data, "aa+v", .verbose = F)
```

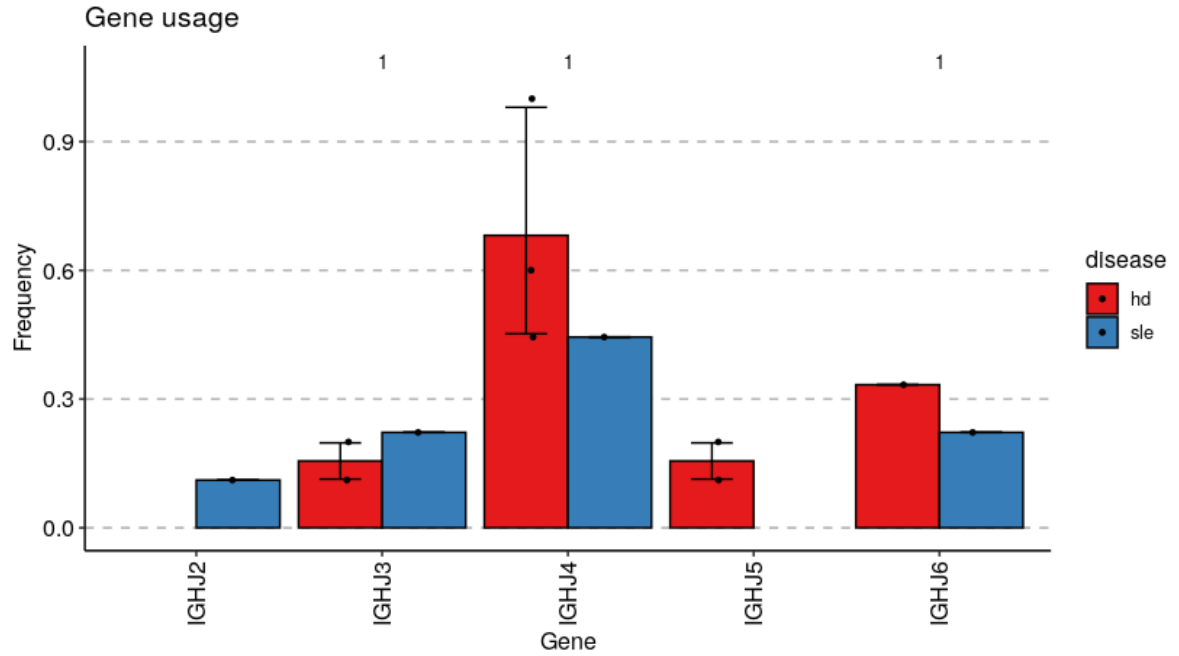
目前可视化了SLE中最为富集的10个克隆型，发现是非常疾病特异的，并且占据了30%以上的比例（一共有超过1000个克隆型）。还可以补充的分析是这个病人在治疗半年之后观察这10个克隆型的变化趋势，mark一下。

这些找出的富集克隆型应该如何研究？（在目前数据库中缺少注释的情况下）>

其中一个克隆型只有轻链而没有重链信息，可能是没有测到匹配的重链？



基因使用偏好性：这部分我还没有完全确定。当cellranger在mapping reads到VDJ基因时不是特别确定，就会同时返回多个可能的基因（因为VDJ免疫基因家族中的复等位基因的序列差异本来就不大）。所以当在计算基因使用偏好时候就有多种策略：1）只使用那些精确mapping的gene； 2）全部都使用(将mapping到多个基因的片段视为不同)。目前采用的是策略1 ，可能需要查一查文献确认一下哪一种方案更好

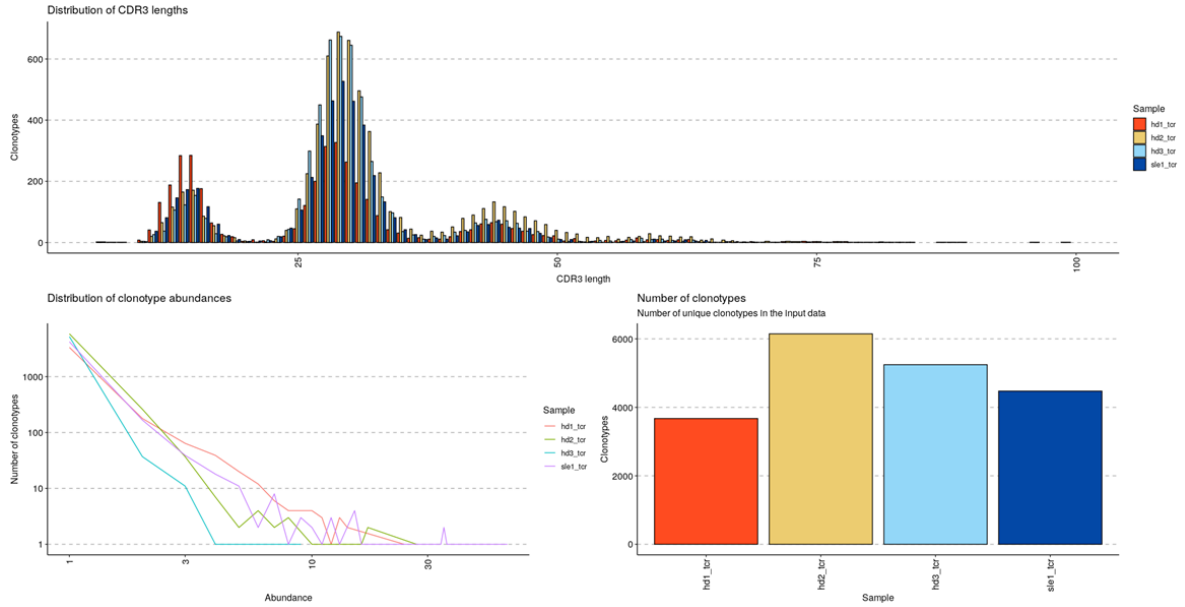


其他分析：

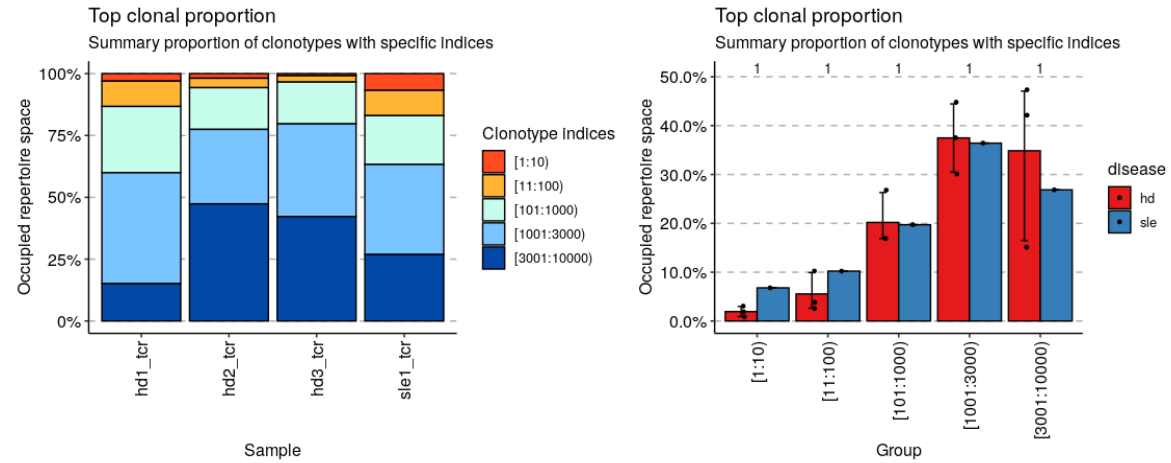
1. 数据库注释：现在BCR数据库TBADB中只有900多条BCR的记录，没有与我们数据中的交集
2. CDR3区域的k-mer和motif分析：暂时没有想到有什么帮助，一般是给机器学习算法用的

4.2 TCR（只保留了最关键的图）

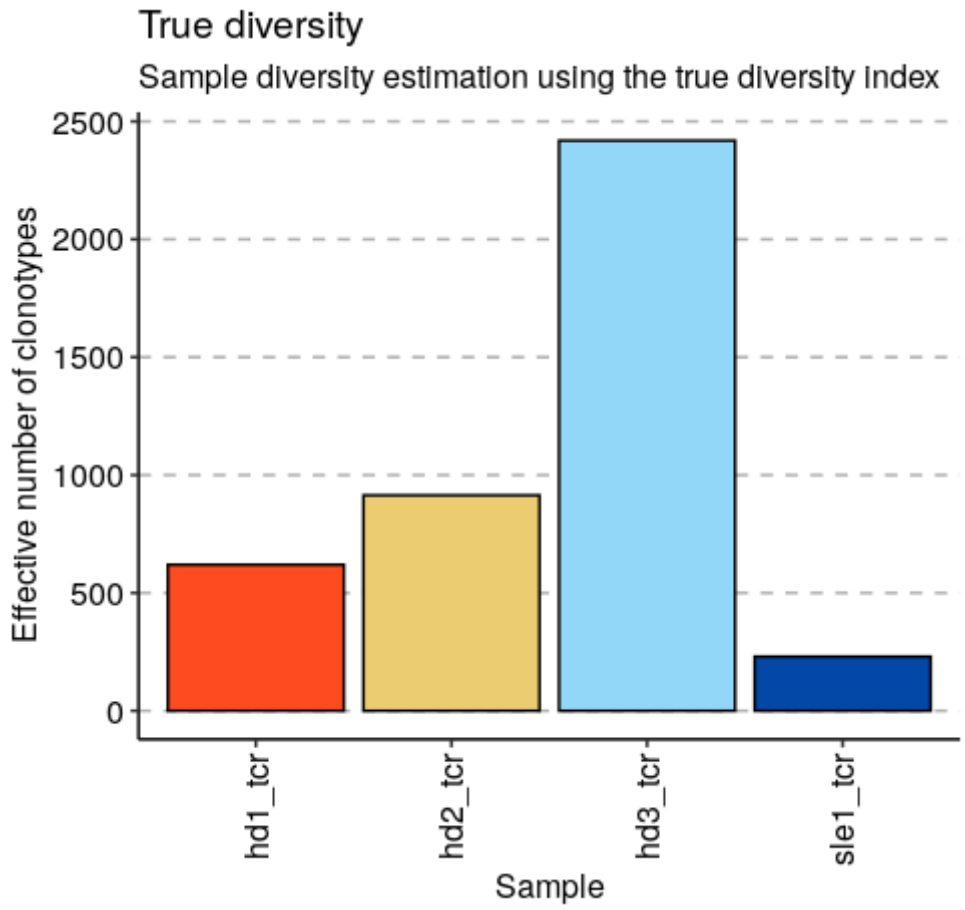
类似BCR，观察到4个样本测到的TCR克隆数目也是差不多的。但是左下图仍然可以发现SLE出现了一个明显的克隆型富集



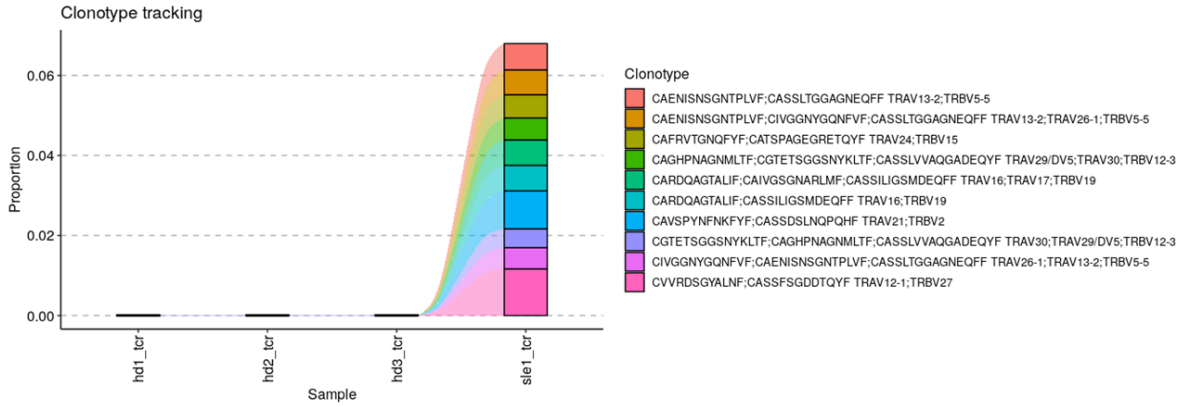
与BCR类似，可以发现TOP10富集的克隆型在SLE中所占比例要高于HD对照（左图是单个样本，右图是分为两组）



用Ture diversity指数回归了克隆型个数的影响吗，发现SLE中的克隆型多样性还是最低的



同理，显示SLE患者TCR组库中最为富集的克隆型



5 一些看法

VDJ是我们的一大优势，因为这是第一次对SLE的immune repertoire 进行系统性描述的工作，可能要集中在以下几个问题

- **SLE中特异富集克隆型与疾病机制的潜在联系**：有两个思路 1）克隆型本身的序列与疾病的关系，目前数据库中的注释不足，所以可能要采取和Cell文章中类似的分子动力学方法（我暂时不会） 2）从拥有这些克隆型的细胞入手，看这些细胞有什么特点（所属的亚类，或者表达特异的marker gene等），但不知道是否可行（下一步工作）
- **这些克隆型在治疗前后的变化**（同一病人）：不同病人之间的异质性很大，且不同人的自身抗原也大不相同，我不对不同病人之间share同一克隆型抱有太大希望。只要能解释清楚疾病治疗对一个人免疫组库的变化就已经非常有意义了，在癌症领域这样的工作都不是很多（因为取样难）