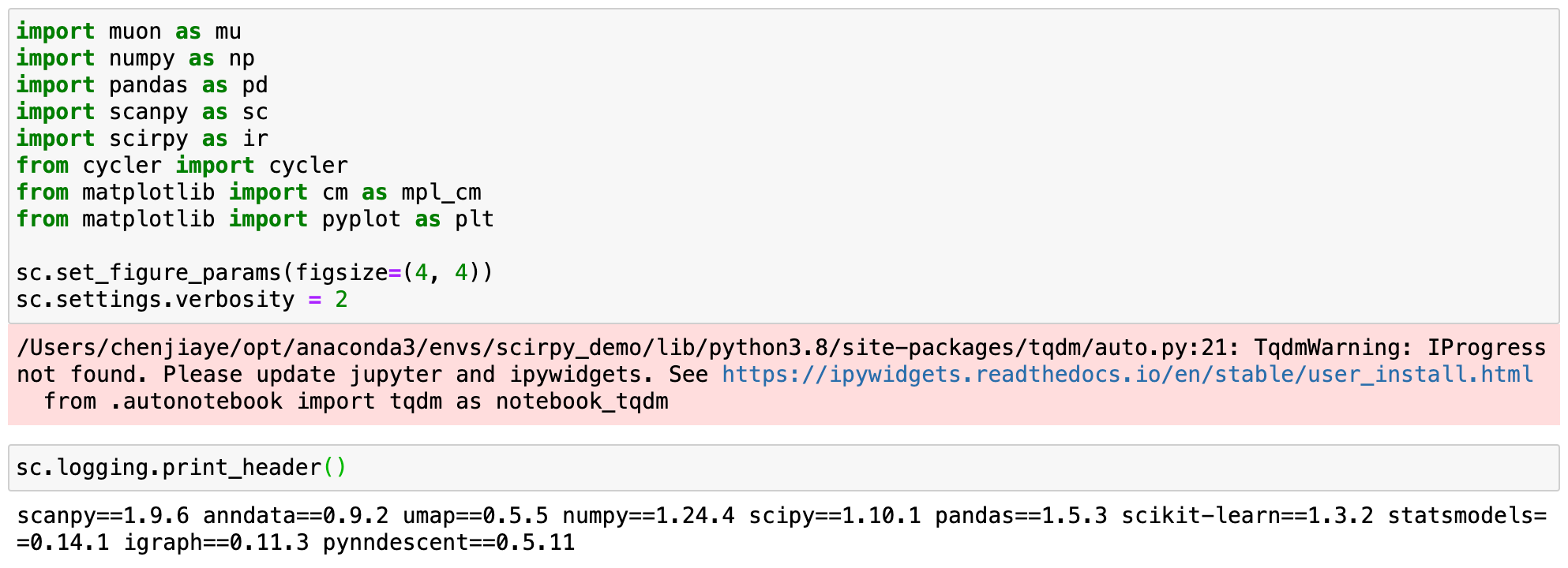
**单细胞免疫组库分析——实例展示（Scirpy）**

在第六章第一节单细胞免疫组库分析中，我们介绍了一款基于Python语言开发的软件工具---Scirpy，并提到了这款软件可以和单细胞转录组数据分析软件Scanpy很好的结合，使得我们可以轻松实现TCR和转录组数据的多项联合分析。这里，我们将使用Scirpy软件内部自带的一套实例数据演示Scirpy软件中能够实现的各项分析内容 (Scirpy Tutorial)。这例数据来自一篇名为《Peripheral T cell expansion predicts tumour infiltration and clinical response》，并于2020年发表在《Nature》上的文章(Wu et al., Nature, 2020)。这项研究的数据中包含了>14万的T细胞，但在Scirpy提供的展示数据中，为了计算效率，我们只包含了3000个细胞。下图中是所有我们分析中需要使用到的python工具以及他们的版本。



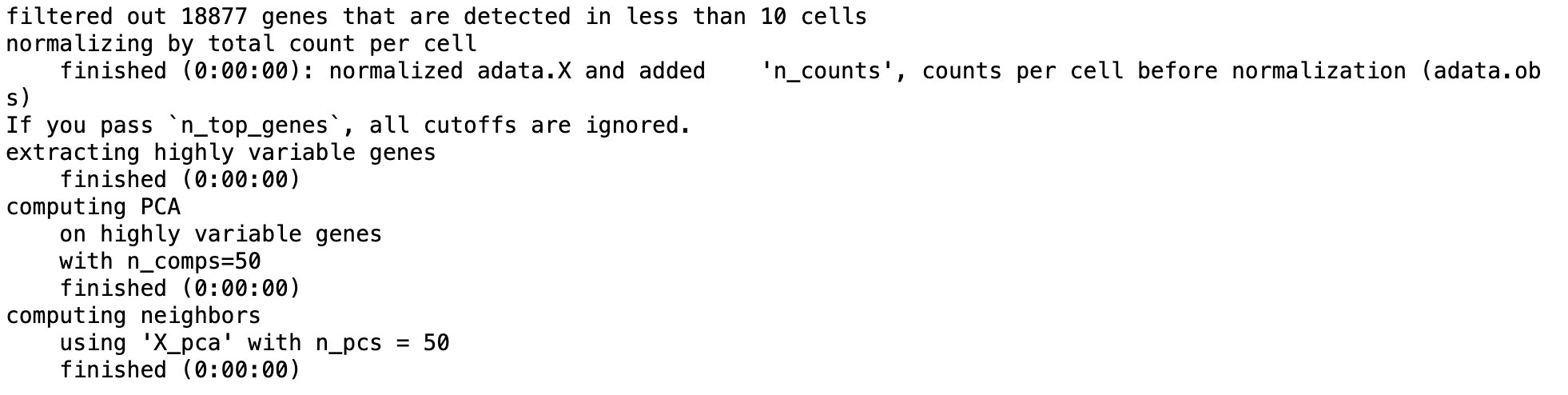
**数据载入以及预处理**

这里，因为来自Wu et al. 的数据已被包含在Scirpy软件包内，我们只需要通过datasets模块录入即可。



此前我们已经提过，Scirpy软件和Scanpy能够很好的结合，因此对于数据的预处理，例如：质控过滤基因和细胞，归一化，标准化，降维聚类等等，我们可以使用常规的Scanpy流程。





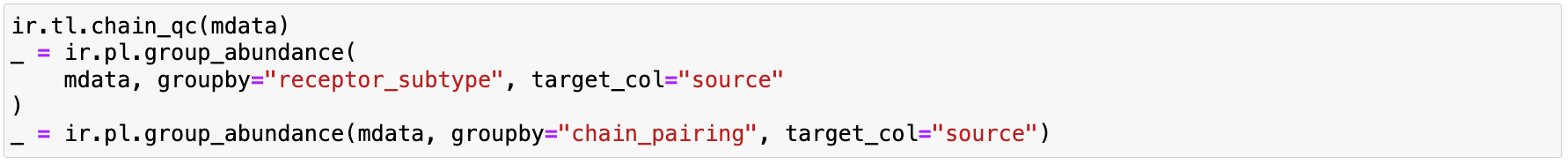
简单的数据与处理过后，我们可以将样本，病人编号等信息通过UMAP图可视化用于辨别数据中是否有批次效应存在。同样，我们也可以将一些已知细胞类型的经典标记基因表达展示在UMAP图中，例如这边展示的CD8A（CD8+ T细胞）和CD4，CD3D/E/G（CD4+ T细胞）。

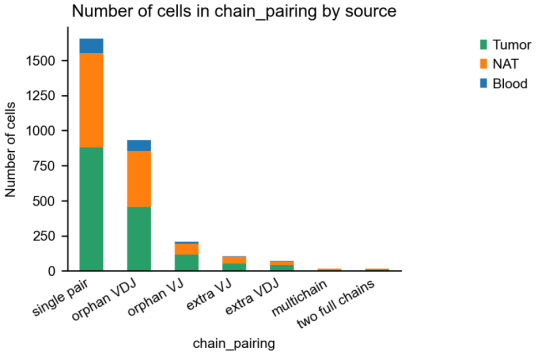
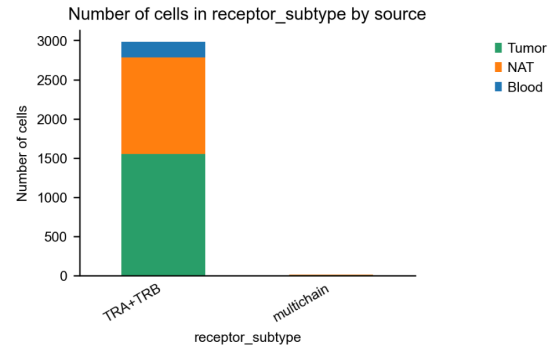




**TCR质控**

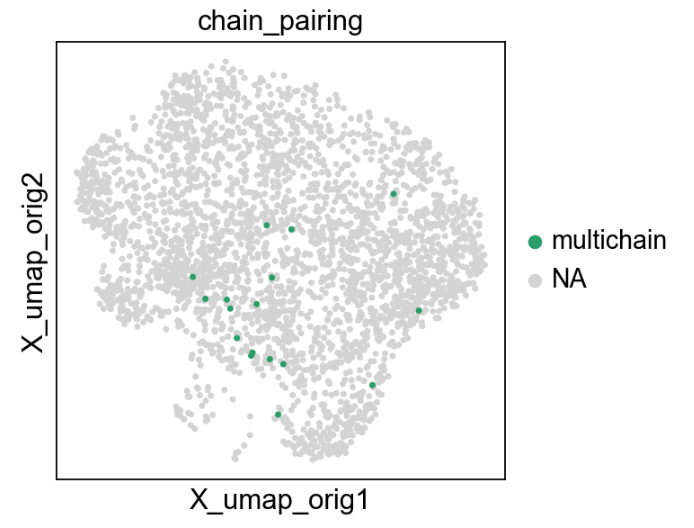
通常，我们知道TCR只含有一对配对的alpha和beta链，但事实上有将近三分之一的TCR可能会存在双TCR（Dual TCR）,即来自不同等位基因的两对受体。一般我们会在下游分析中将其去除。





这里，我们可以通过UMAP图将拥有多条链的T细胞可视化出来，并在后续的分析中去除这部分细胞。同时我们也可以将那些没有完整的一对链的T细胞也一起去除。



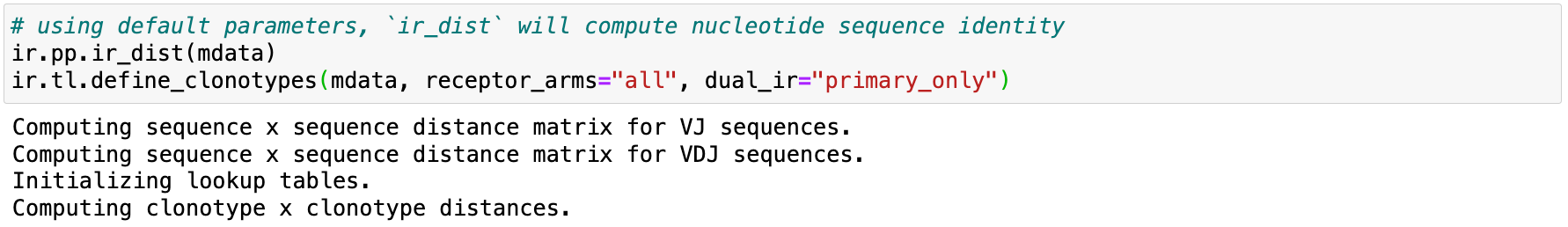




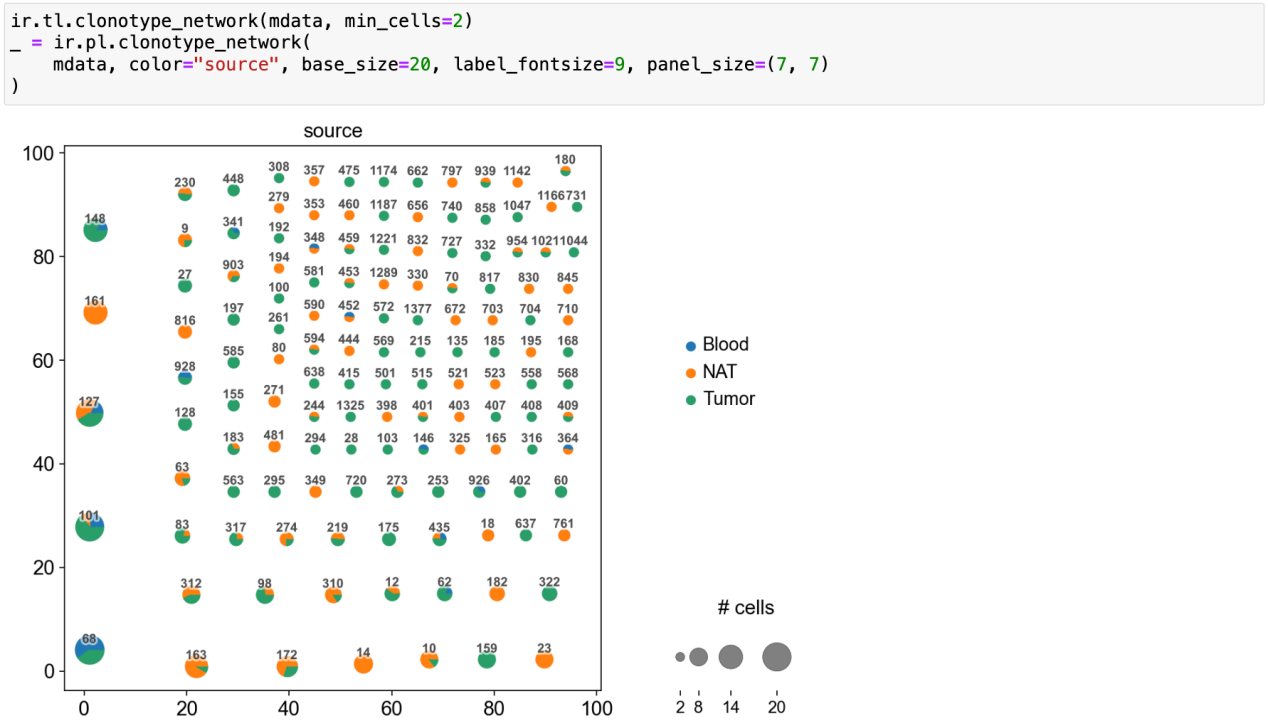
**定义克隆型以及克隆型聚类**

Scirpy中计算和定义克隆型使用到了一种基于网络（network-based）的算法。这种算法类似于处理单细胞转录组数据时，在聚类之前创建领域图（neighborhood graph）的过程。

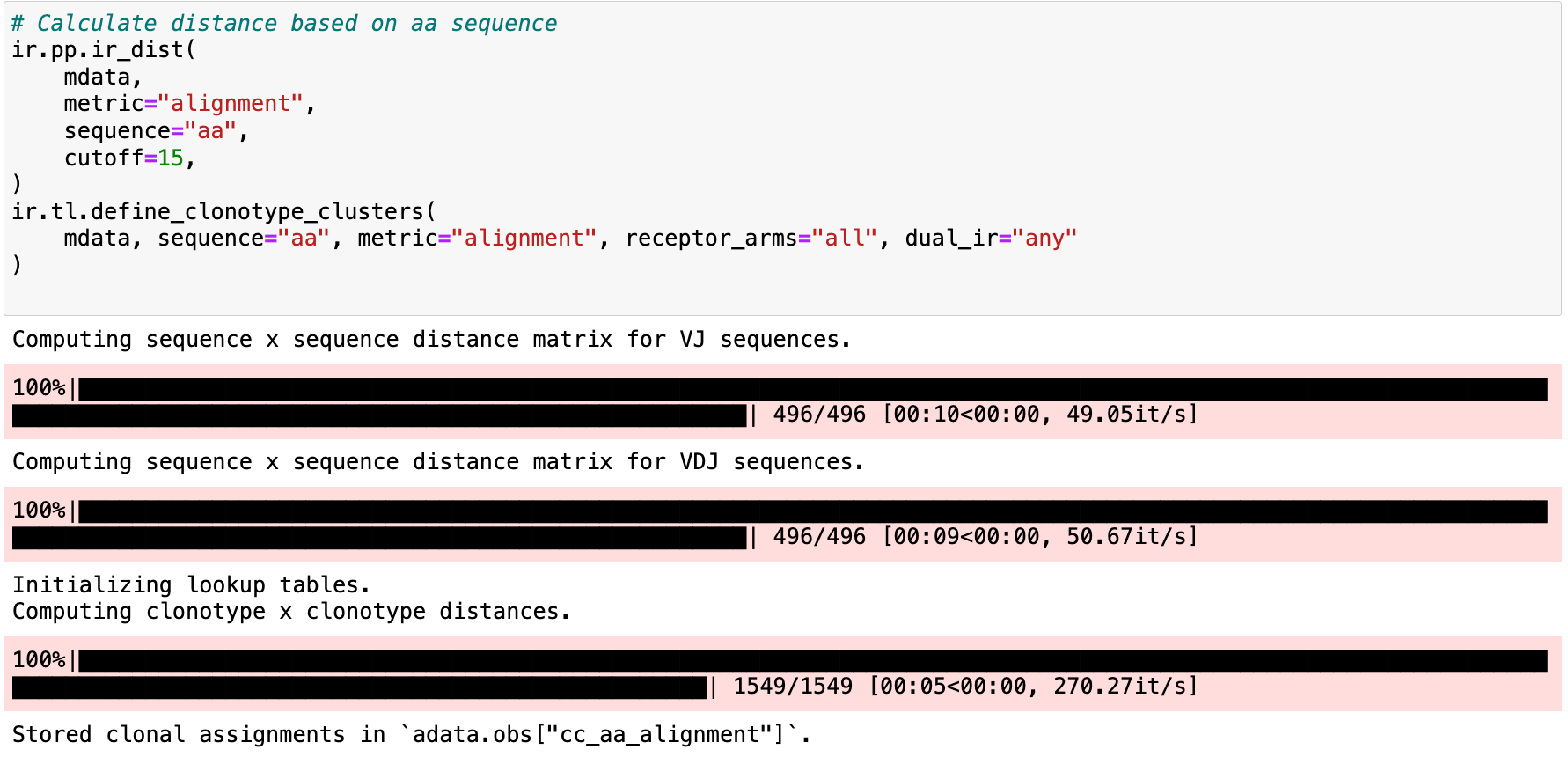
这里，我们使用到的ir.pp.ir\_dist函数实际上是在计算CDR3序列之间的距离。这里的序列可以是RNA对应碱基序列（nt），也可以是蛋白质对于的氨基酸序列（aa），可以通过改变sequence参数进行选择。我们在下图中看到的是用默认参数，也就是基于碱基序列计算得到的结果。

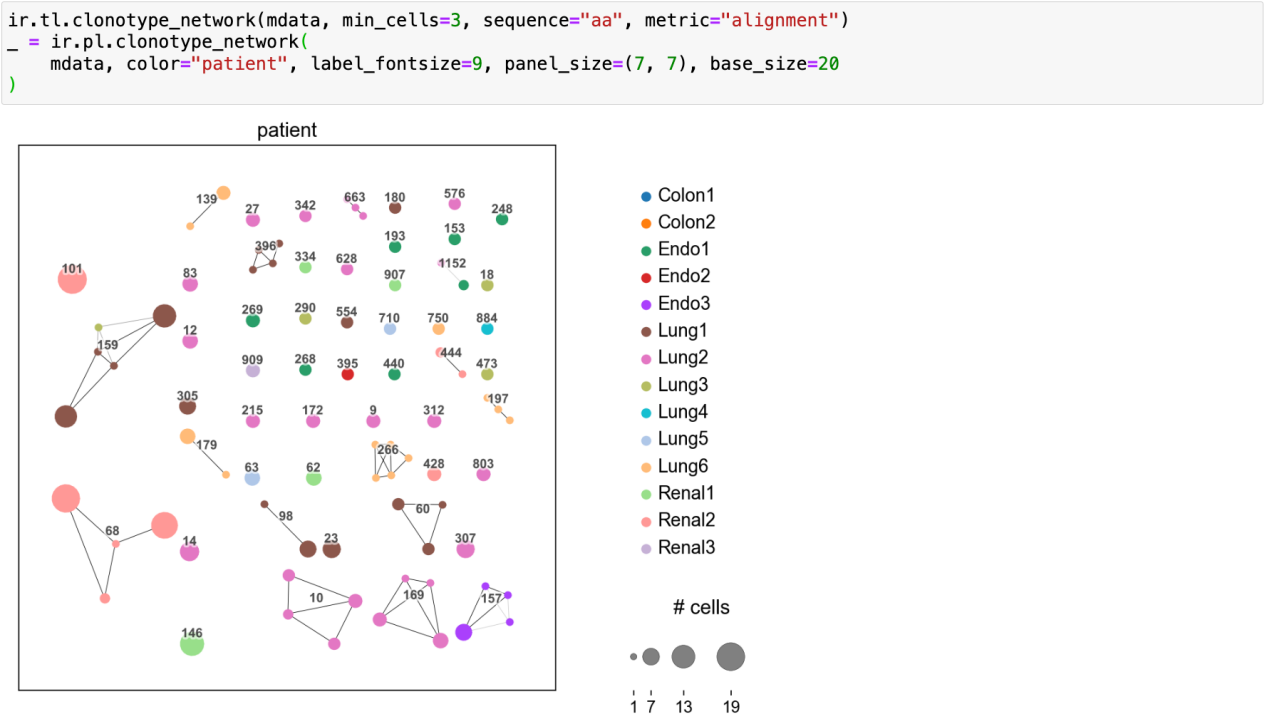


在可视化的时候，我们可以将每个克隆型内部包含的最小细胞数阈值设置成2 。这样，我们可以在可视化时避免看到单一细胞构成的克隆型。



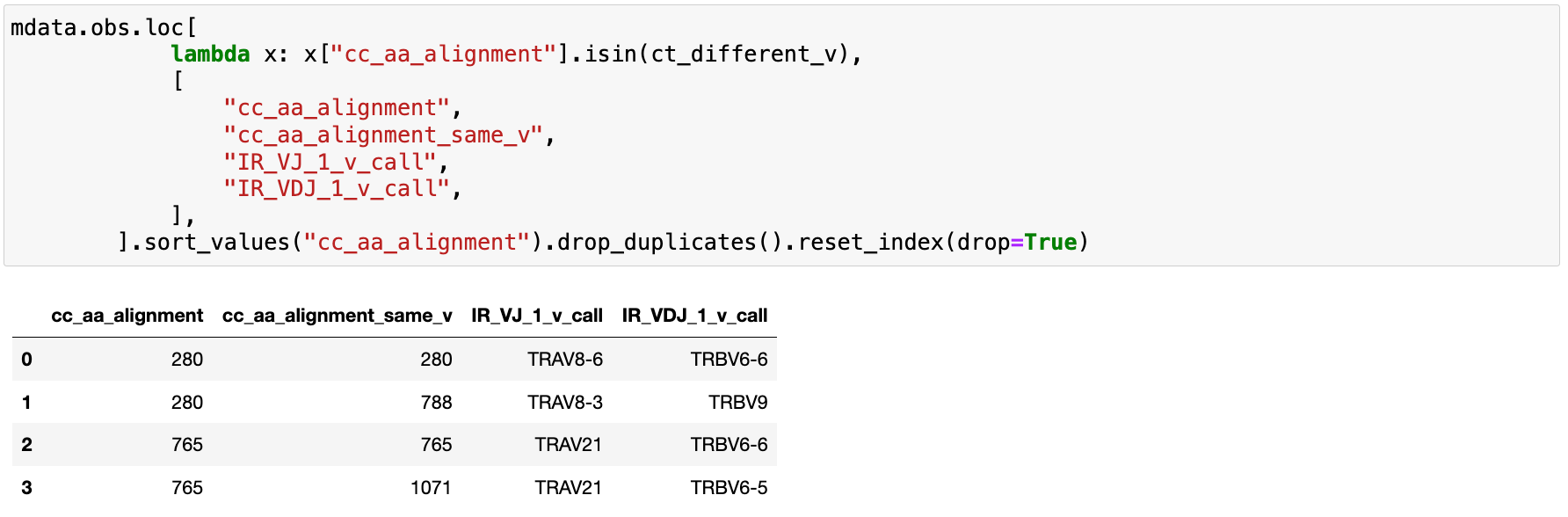
接下来，我们可以尝试将sequence参数改设成 “aa”使得我们可以计算氨基酸序列间的相似性。这里，我们需要同时将metric参数改成“alignment”，因为Scirpy中计算氨基酸序列的相似性会用到基于BLOSUM62矩阵的比对算法。这里的cutoff参数决定了在计算完相似性之后，分值小于多少的相邻克隆型会被连接在一起形成一个网络。



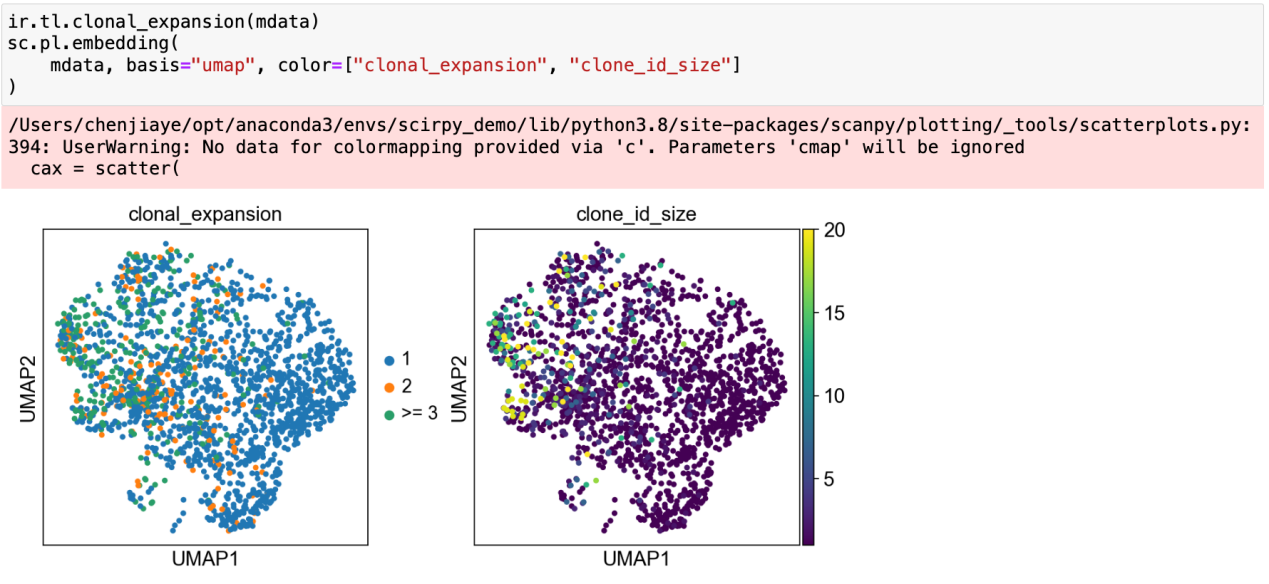


这里，在定义克隆型的过程中我们还可以添加一个“same\_v\_gene”参数。当这个参数被设置为True时，会使得在定义克隆型时，被定义为同一克隆型的所有细胞都具有相同的V基因，既是相同的CDR1,2区域。在下图中我们可以看到，原本被归为280和765的两个克隆型，因为same\_v\_gene的设置，被分别重新拆分成两个新的克隆型280 （280，788）；765（765，1071）。

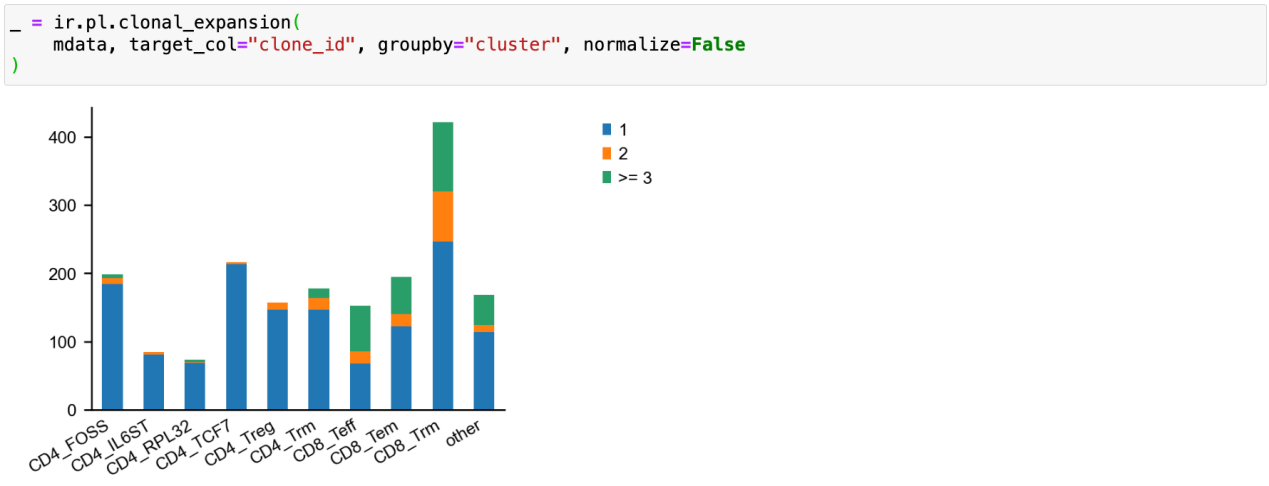




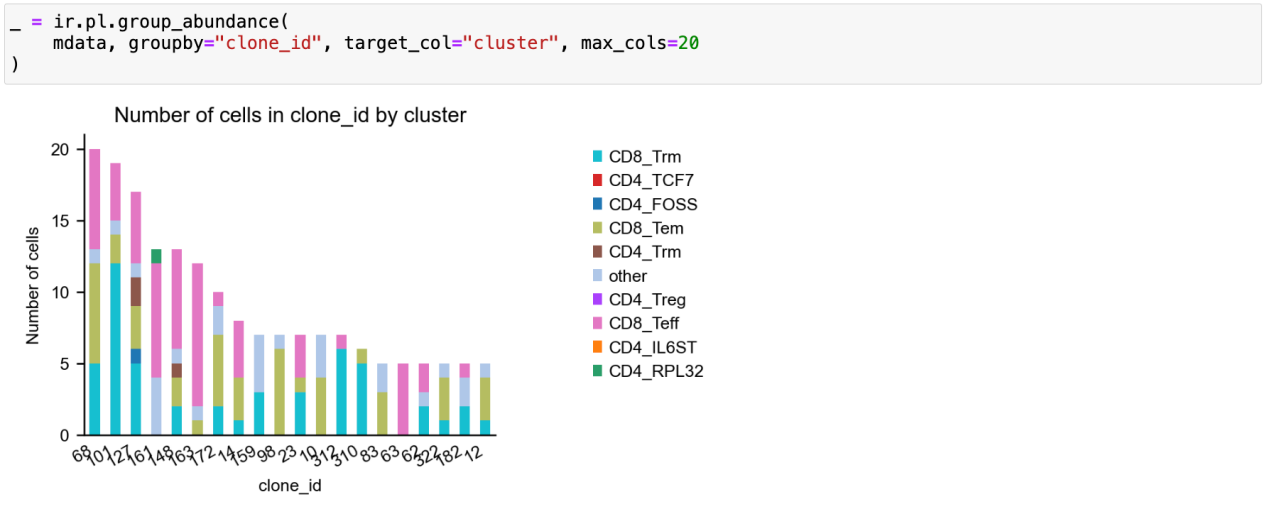
在定义完克隆型后，我们可以通过UMAP图可视化那些包含了两个及以上细胞的克隆型（Expanded Clonotypes）。



同时，也可以通过柱状图展示每个不同细胞类型中，expanded clonotype的数量。

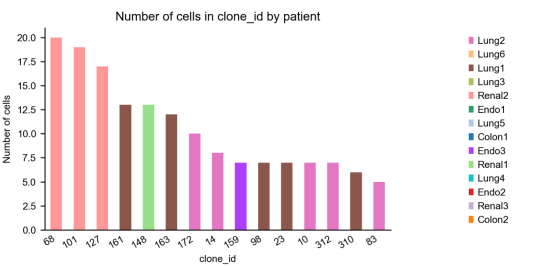
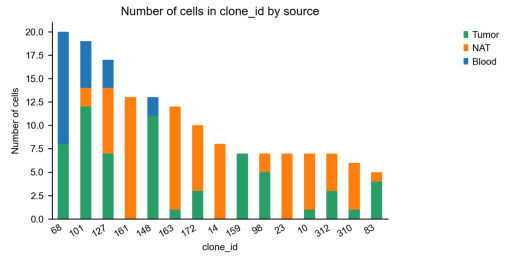


反过来，我们也可以展示在不同克隆型当中，不同细胞类型的数量分布。

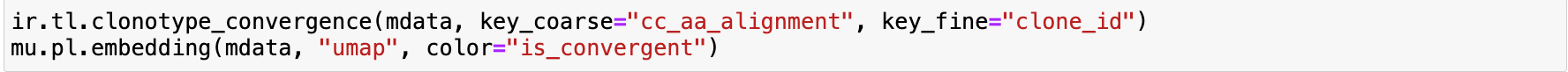


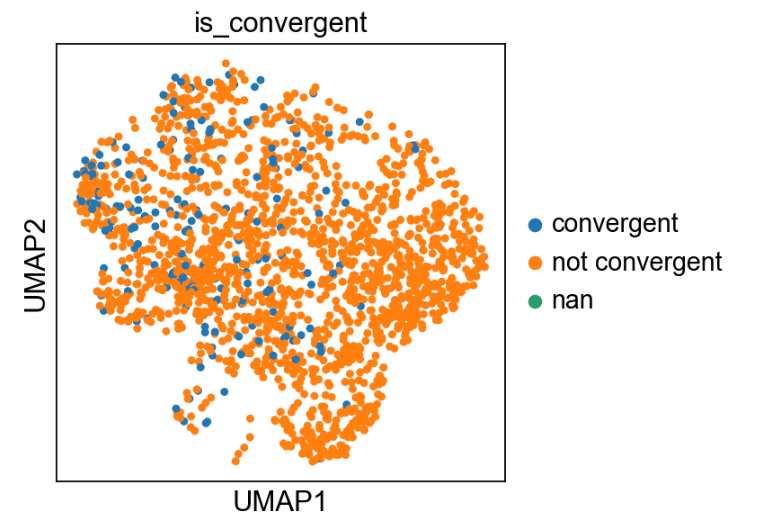
我们还可以展示不同克隆型中的各种样本类型的分布，比如说是否为癌症样本，或者病人的编号。这样，我们可以看出哪一些克隆型是不同样本之间共有的，而哪一些是只富集在某一位病人的样本中或是所有癌症病人中的。





最后，我们还可以通过克隆型看出一些进化的规律。我们可以去看哪些克隆型虽然序列不同，但是却能够识别同一种抗原。我们认为这种进化类型是趋同的。

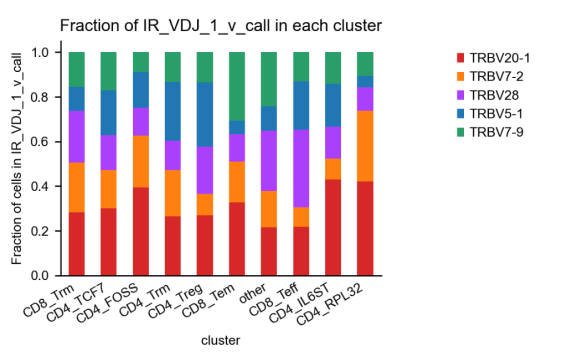
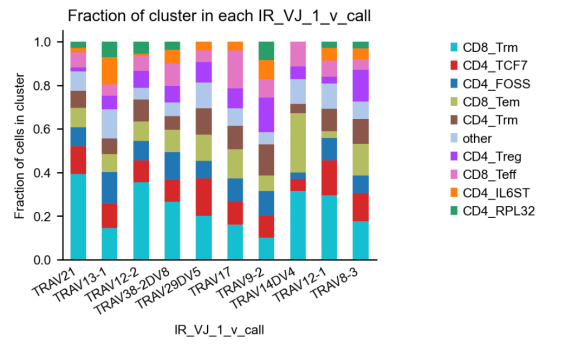




**基因的使用**

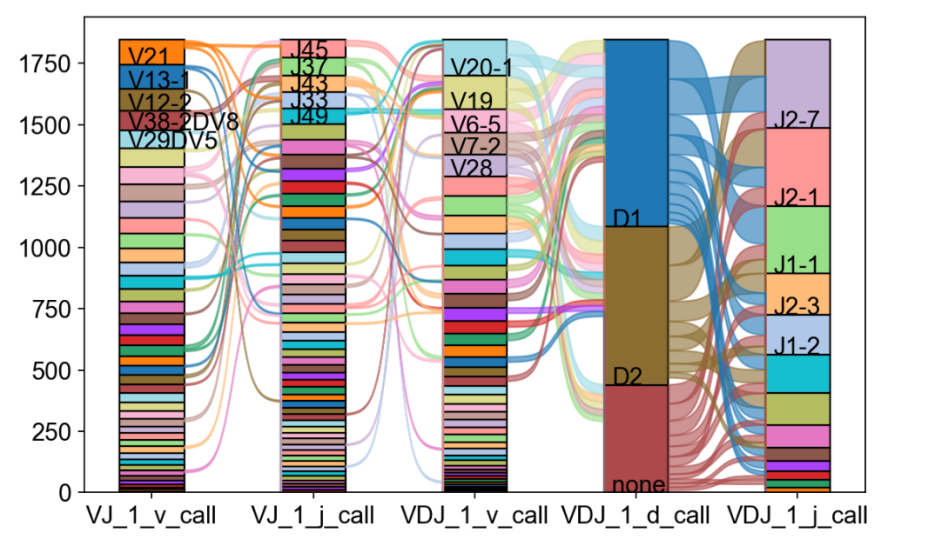
探究完克隆型本身的一些信息，我们还可以从免疫组库中获取这些克隆型具体的VDJ基因使用情况。比如在下图中，我们可以可视化V基因和细胞类型的关联。

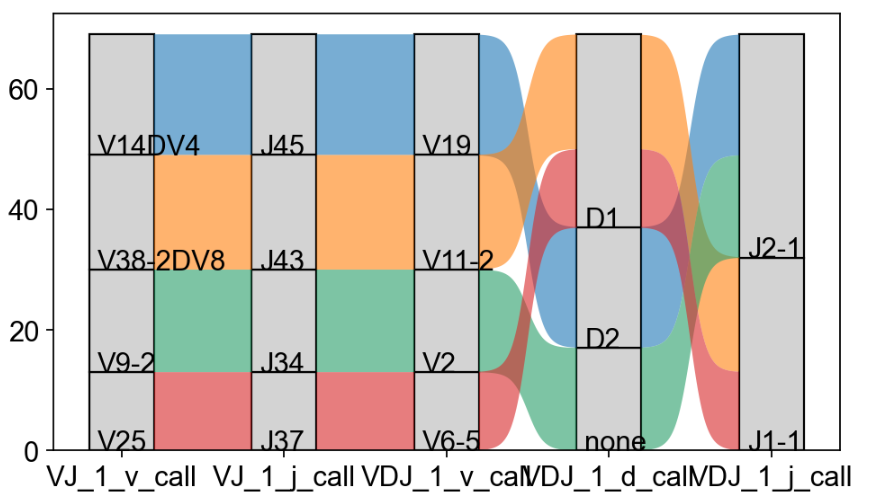




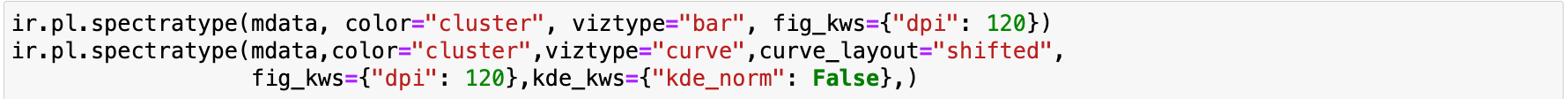
我们也可以通过Scirpy.pl.vdj\_usage函数可视化每个克隆型或是具体几个克隆型具体的VDJ使用情况。

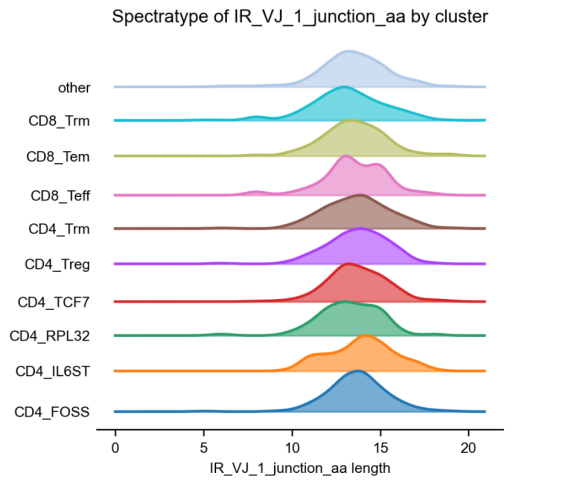
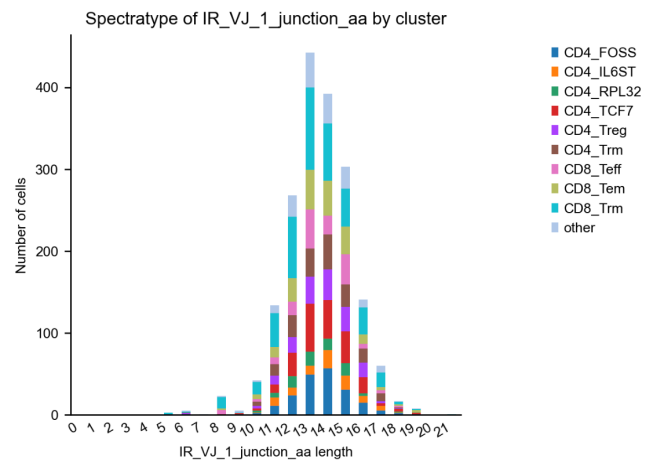






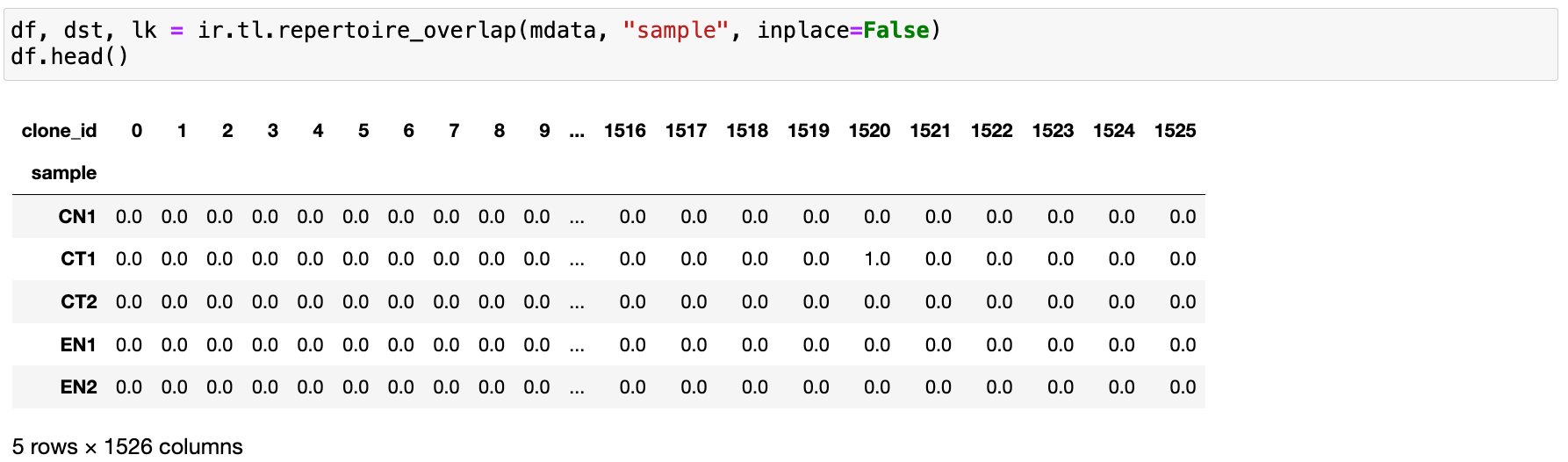
同时，通过spectratype函数，我们可以可视化CDR3区域长度在不同细胞类型中的分布。通过柱状图可以看出具体的占比，而山峦图可以很好的展现CDR3区域长度的分布情况。

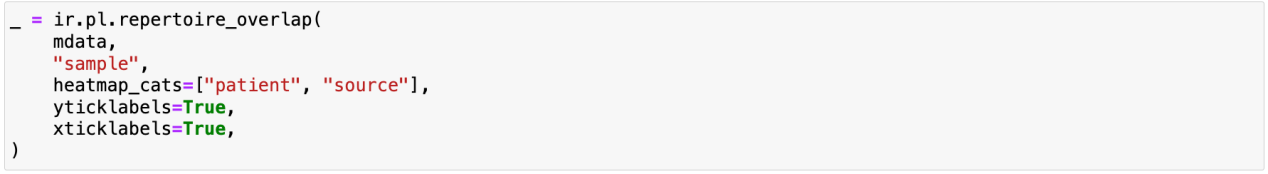


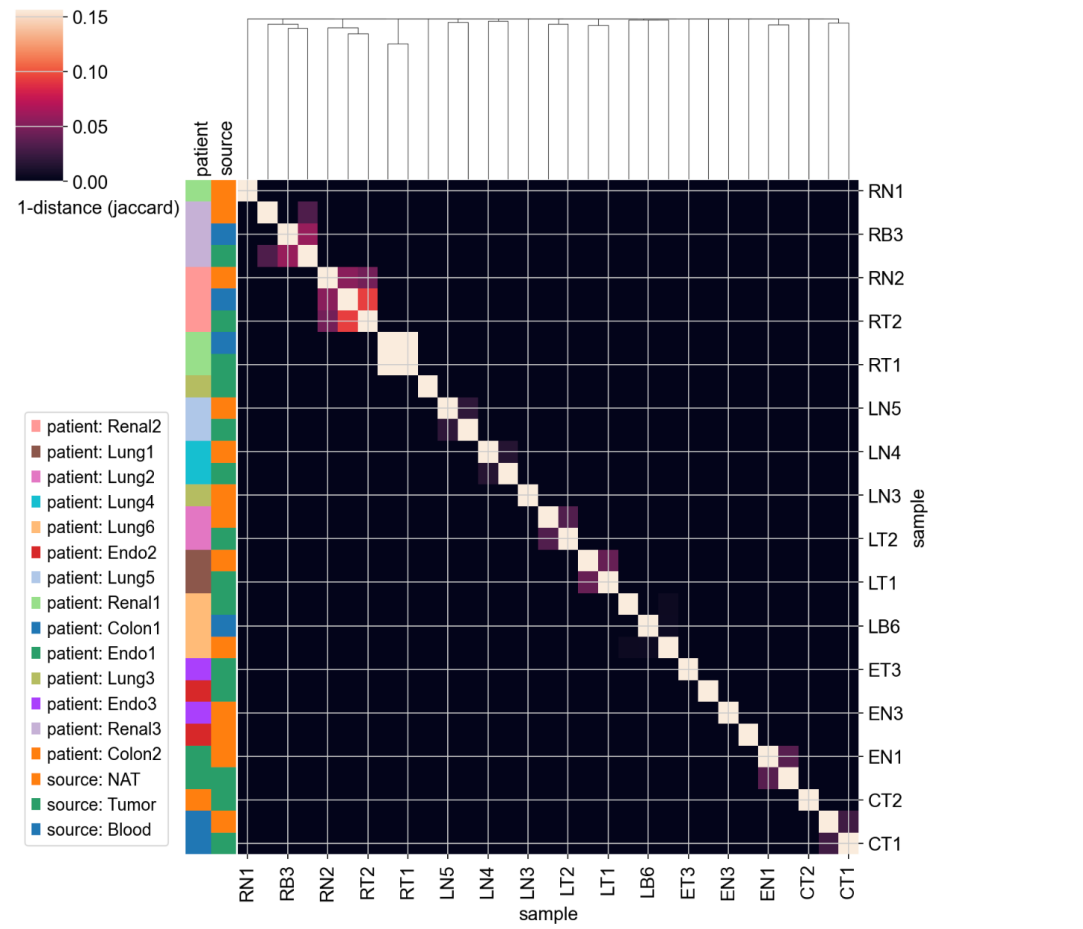


**免疫组库对比**

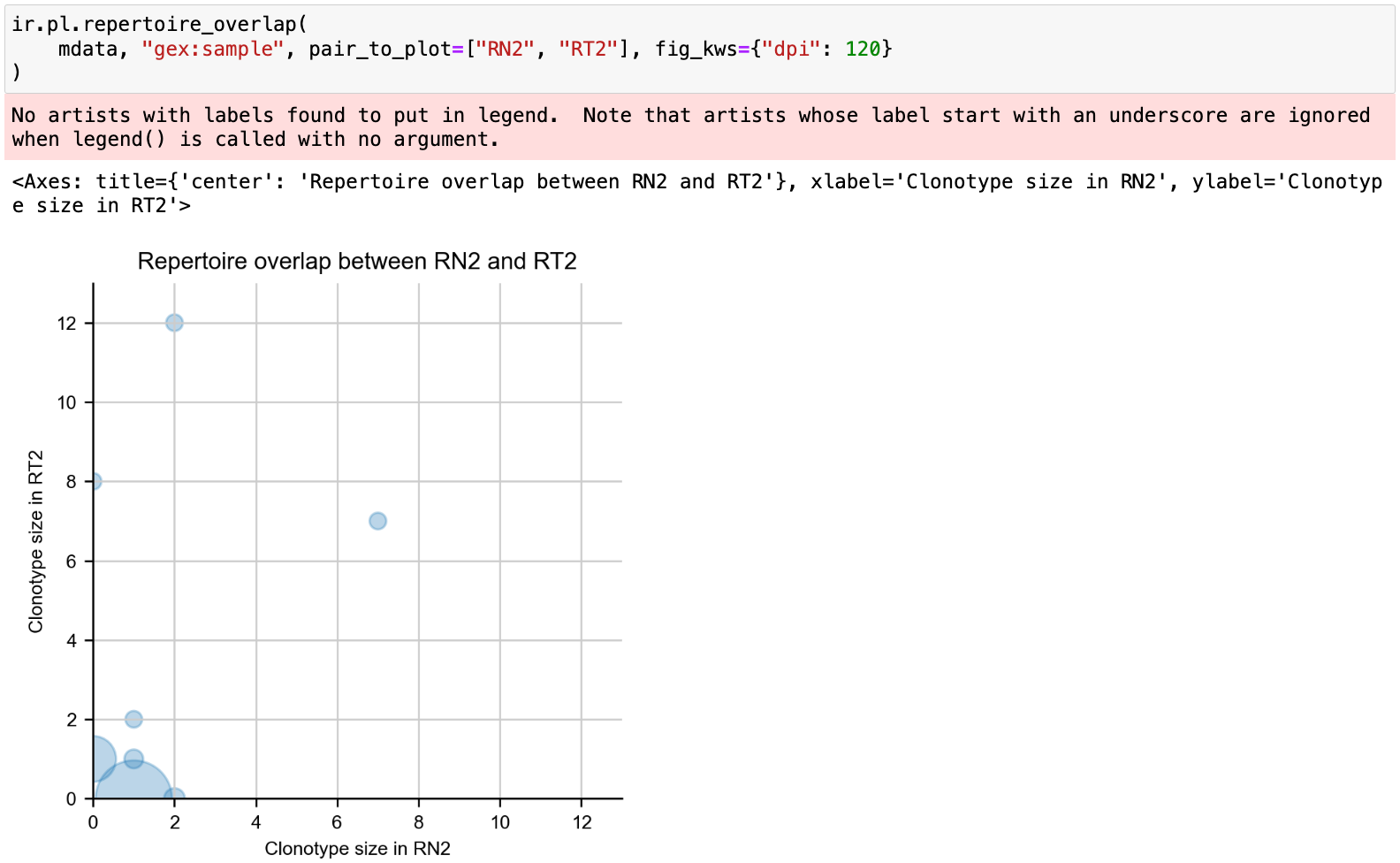
在免疫组库分析中，一些比较重要的克隆型往往不会只在一个细胞类型，或是一个样本当中出现。为了去发现那些在不同样本中都出现了的克隆型，去计算样本间免疫组库的相似性就是必要的。这里，Scirpy中提供了repertoire\_overlap函数去计算不同克隆型在不同样本中的丰度。运行结束后，我们会得到一个样本乘克隆型的矩阵，矩阵内部的数值则代表每个克隆型在各个样本中出现的数量。同时，还会产生一个样本间通过Jaccard 相似性计算出来的间距矩阵（Distance Matrix）。我们可以通过热图的形式可视化间距矩阵来看出哪些样本之间的相似度较高。





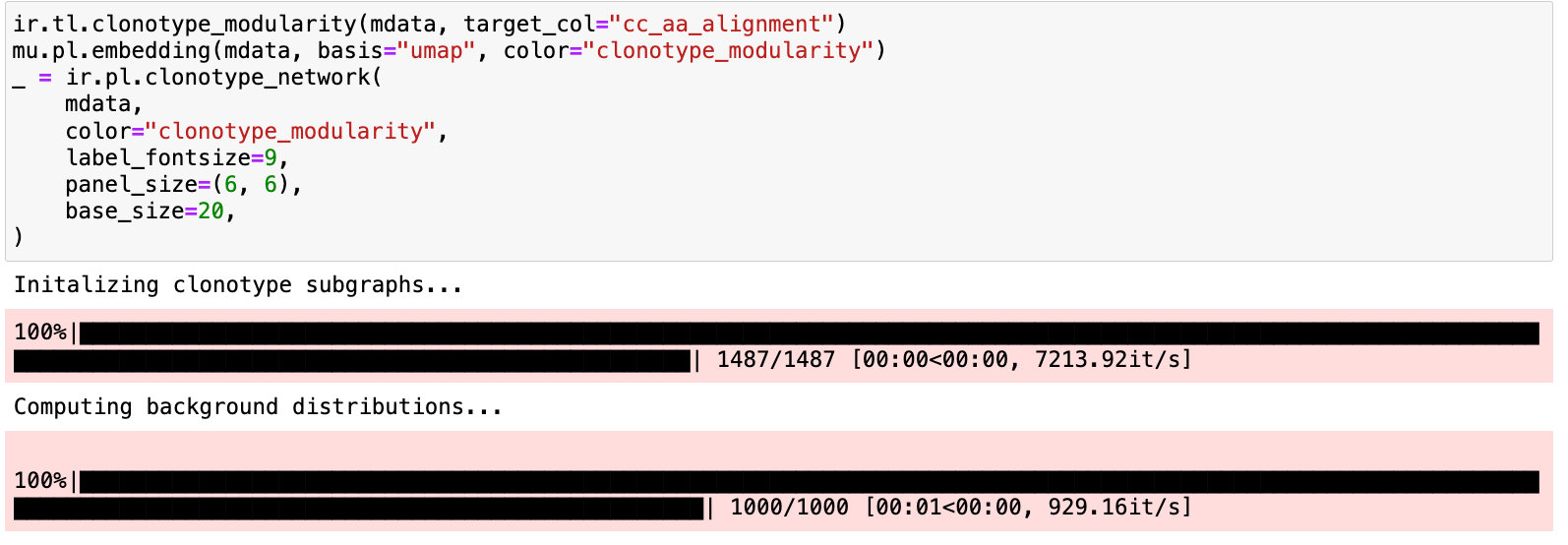


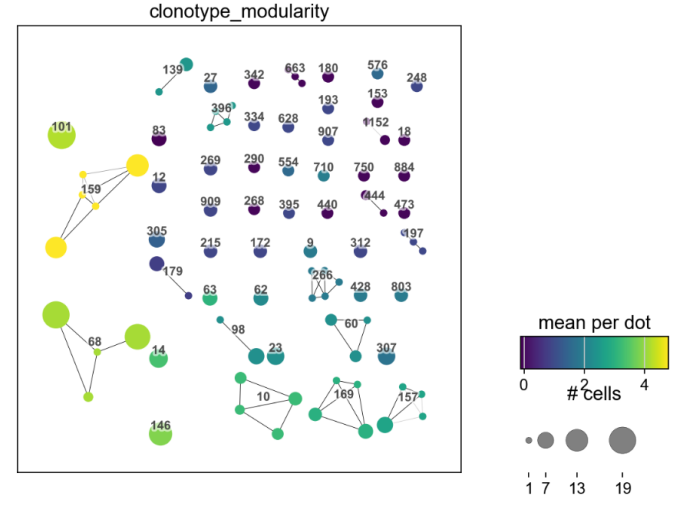
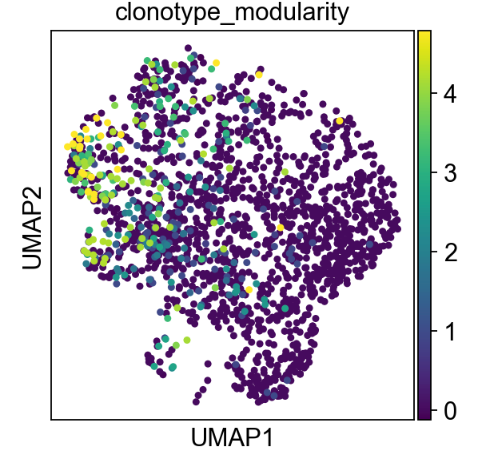
通过热图找到相似度较高的一组样本后，我们还可以单独比较两个样本间重叠克隆型的数量。



**整合基因表达数据**

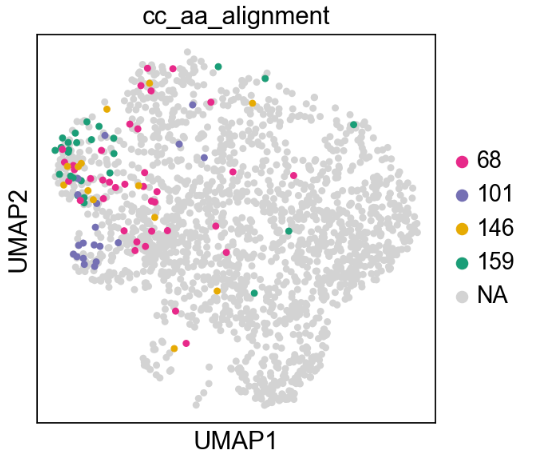
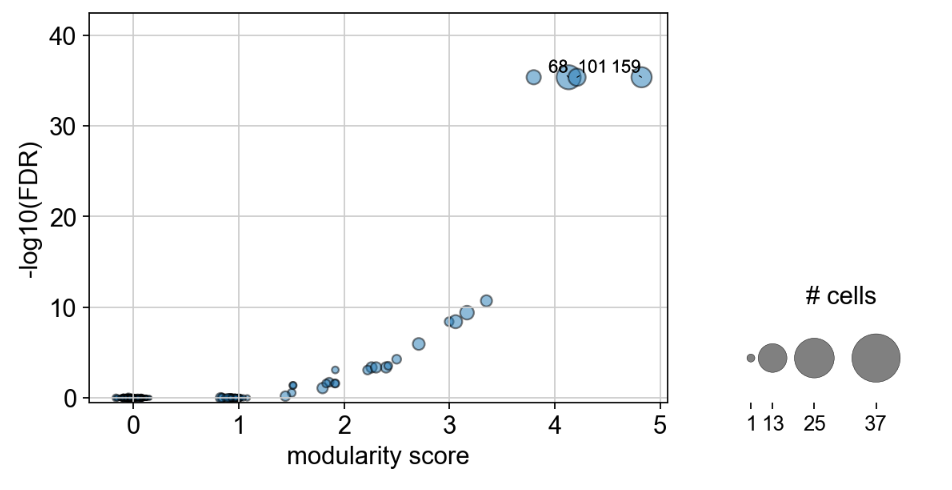
由于Scirpy软件在设计上与Scanpy有非常好的结合度，我们可以很轻松的将免疫组库的数据与转录组的数据合并联合分析。首先，我们可以通过检查同一克隆型当中的细胞是否有非常相似的基因表达模式，从中我们可以得到每个克隆型的模块化指标。这个指标是通过计算一个克隆型中的细胞在细胞领域图中拥有的边（edge）的数量和完全随机的领域图中那些细胞拥有的边的数量的比值获得的。计算出的克隆型模块化分数可以通过UMAP图展示。



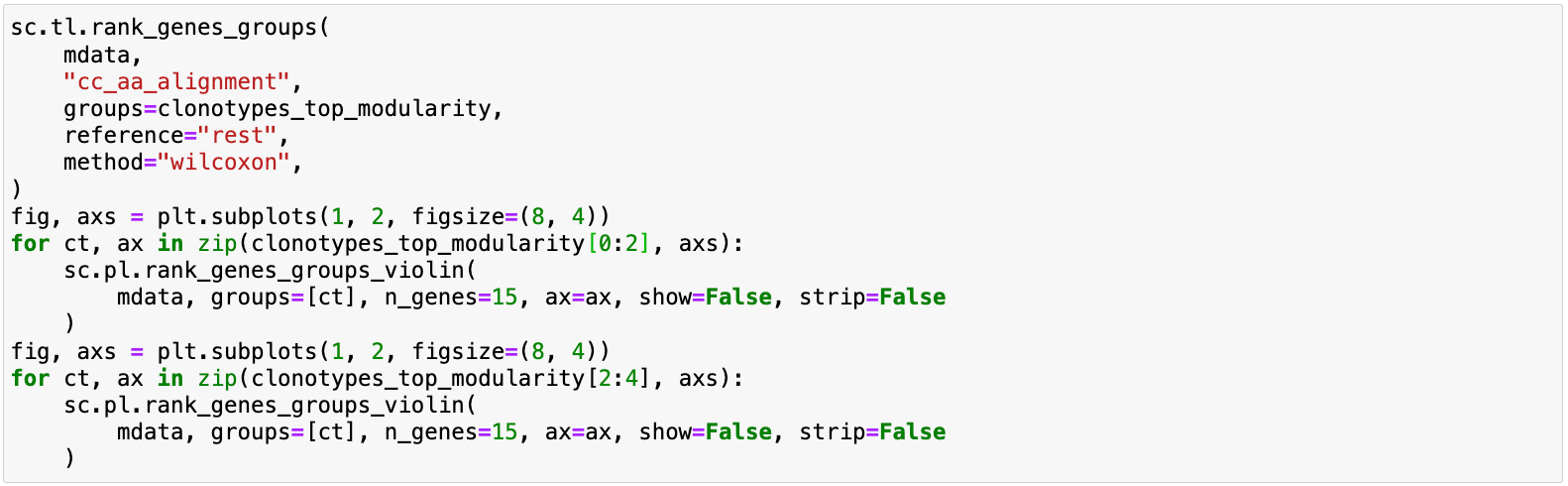


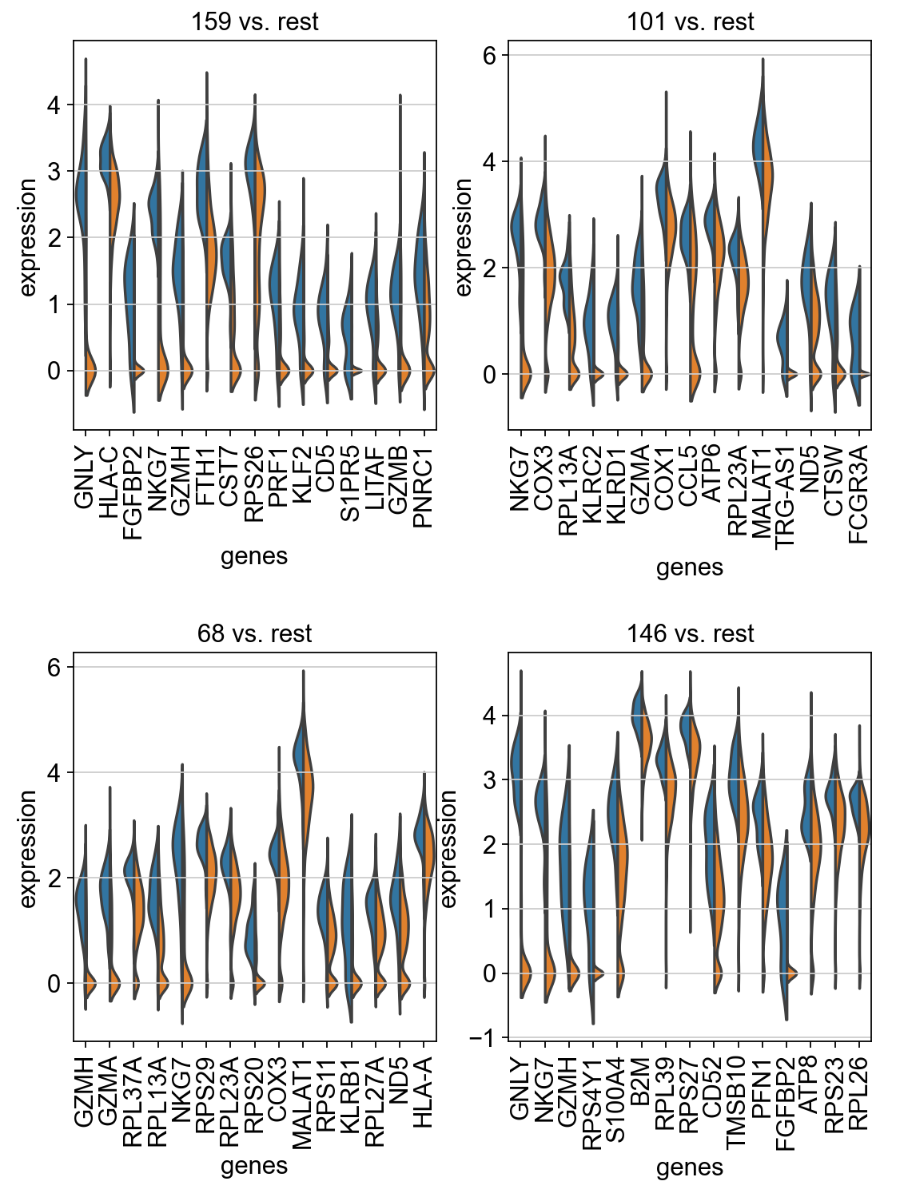
随后，我们可以通过火山图看到有几个克隆型拥有远高于别的克隆型的模块化分数。我们可以将它们挑出来单独展示在UMAP图，看他们分别富集在哪些位置。



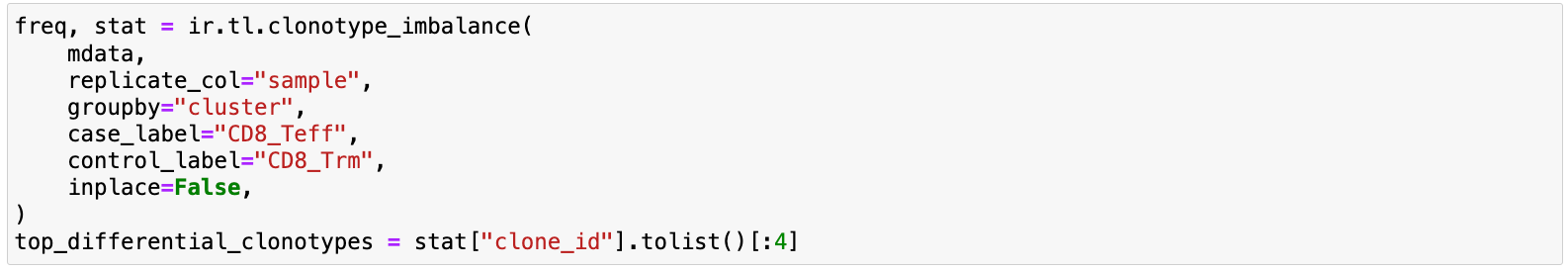


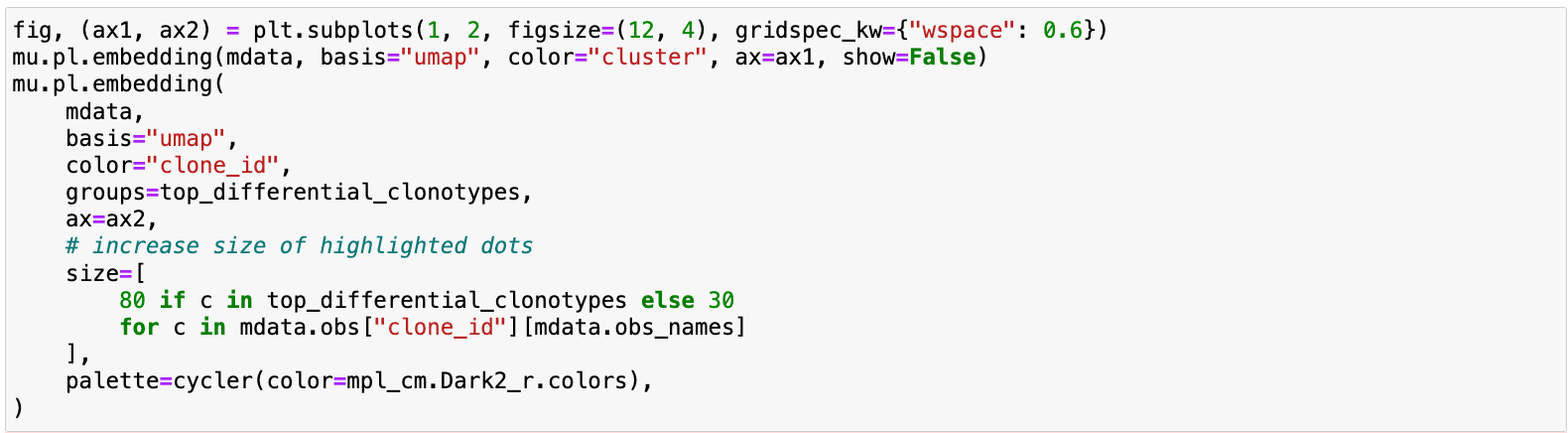
接着，我们可以像计算差异基因一样，计算每个克隆型和其余克隆型之间的差异化表达的基因。这里，我们选取了模块化分数最高的四个克隆型159，101，68，146。

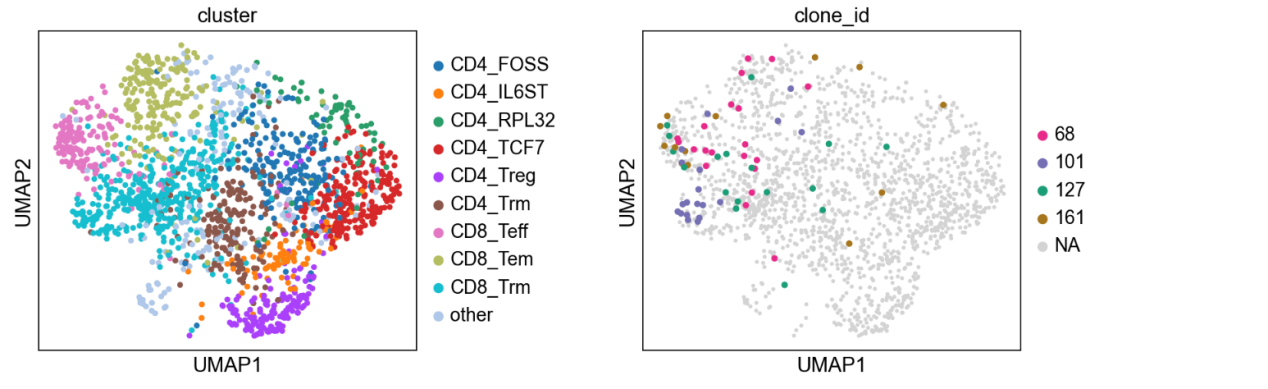




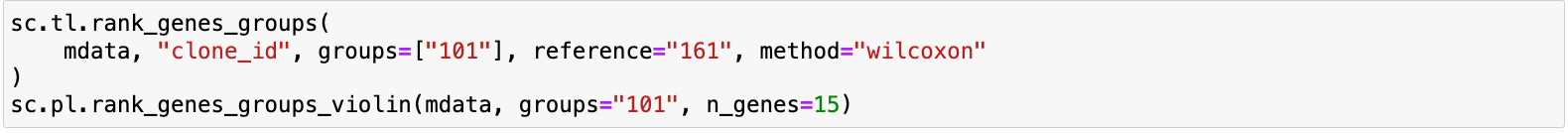
在之前的UMAP图中，我们可以看出来，一些克隆型中富集的细胞是具有一定的细胞类型特异性的。为了更准确的探究这些top克隆型是否有着细胞类型特异性，我们可以用clonotype\_imbalance函数来计算两种细胞类型之间存在丰度差异的克隆型是哪些。这里，我们使用CD8+ tissue-resident memory T cells和CD8+ effector and effector memory cells作为对比的两种细胞类型。

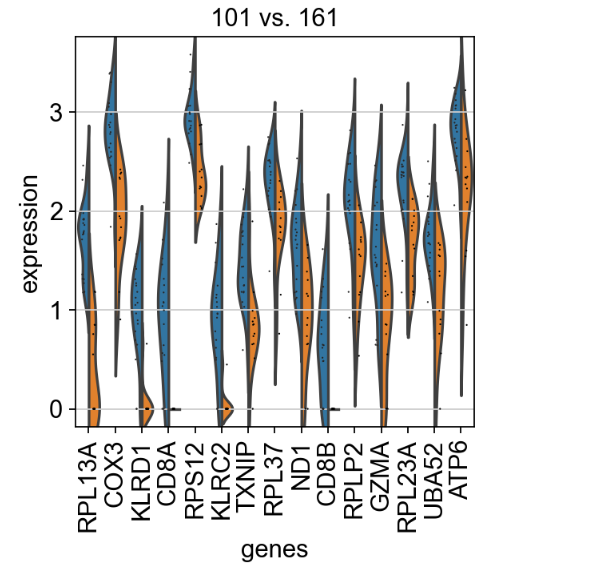






从图中我们可以看出来，克隆型101和161分别富集在CD8+ tissue-resident memory T cells和CD8+ effector and effector memory cells这两种细胞类型中。因此我们可以去计算这两个克隆型中细胞之间差异表达的基因作为这两种克隆型的标记基因。





**使用表位数据库进行免疫受体功能注释**

Scirpy中还有一项重要的功能就是可以通过内部的表位数据库对我们数据中包含的免疫受体进行功能的注释，使我们可以知道该受体是否特异性对应某一种抗原。



