



# 免疫组库测序 结题报告

广州华银医学检验中心

2019年1月8日

## Table of Contents

Table of Contents	1
1 总体工作流程概述	2
1.1 项目介绍	2
1.2 实验流程	2
1.3 生物信息分析流程	2
2 分析结果	4
2.1 数据概况	4
2.2 克隆的鉴定与计数	5
2.3 不同V/D/J基因使用频率	5
2.4 不同V/D/J基因组合频率	6
2.5 CDR3的长度分布	6
2.6 丰富度和多样性评估	7
2.7 样本间相关性分析	8
2.7.1 样本间CDR3序列的相似性	9
2.7.2 样本间VDJ基因以及相互组合的共线性	9
2.7.3 样本间差异假设检验分析	12
2.7.4 样本间PCA分析	12
3 参考文献	13
4 备注	14
4.1 结果文件说明	14
4.2 分析软件说明	14

# 1 总体工作流程概述

免疫组库（Immune Repertoire，IR）是指某个个体在任何特定时间点其循环系统中所有功能多样性B淋巴细胞和T淋巴细胞的总和。T细胞和B细胞分别介导机体的细胞免疫和体液免疫应答，分别通过其表面的T细胞受体（TCR）和 B细胞受体（BCR）来识别和结合抗原，进而发挥功能清除病原体或体内肿瘤细胞。

免疫组库高通量测序是通过多重PCR或SMART RACE法扩增TCR/BCR全长或CDR3区序列（主要是TCR的β链或BCR的重链），再结合高通量测序技术，对机体免疫组库的多样性及每种 T、B细胞克隆的独特性序列组成和变化进行分析，从而全面评估机体的免疫状态，明确疾病与T、B细胞克隆组成及变化之间的关系。随着免疫组库高通量测序技术的不断发展和成熟，基于免疫组库多样性变化特点的生物标志物发现、肿瘤等疾病疗效预测、疾病的易感性和抵抗性、感染性疾病及疫苗研究等方面都取得了重要进展。

## 1.1 项目介绍

项目编号	HUAYINNGS20181206
分析类型	TCRB
比较组信息	WAS9CD8TCM,WAS6CD8TCM,WAS4CD8TCM,WAS8CD8TCM VS WAS4CD4TEM,WAS6CD4TEM,WAS8CD4TEM,WAS9CD4TEM

## 1.2 实验流程

接收样品后，首先对样品进行质量检测；然后提取 RNA，用 添加UMI改良的5'RACE 法扩增免疫组库序列，回收目的片段，构建测序文库；并分别用 Qubit、Agilent 和 Q-PCR 法对文库的浓度、片段的完整性及插入片段大小、文库有效浓度进行检测和精确定量；检测合格的文库用 Illumina 高通量测序平台（HiSeq/MiSeq）进行测序。

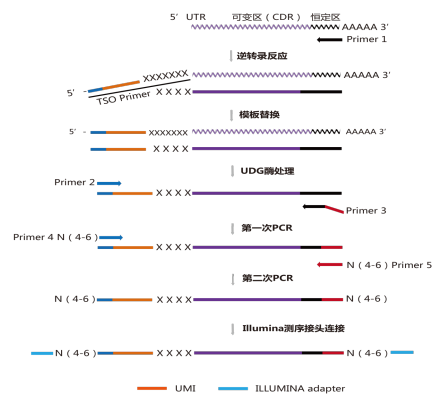


图 1-1 实验流程图

## 1.3 生物信息分析流程

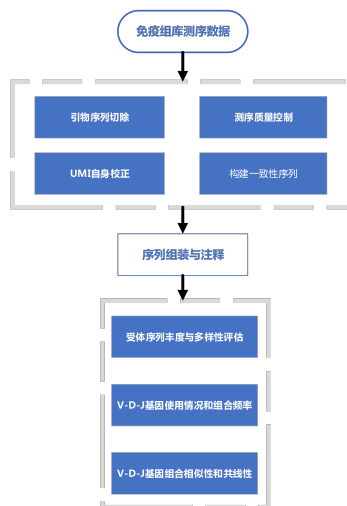


图 1-2 生物信息分析流程图

信息分析分以下几个步骤：

**1 原始下机数据处理** 此步骤过滤测序质量值低的reads，保留高质量reads，过滤后的数据称为Clean Data；

**2 UMI自身校正和根据UMI构建一致性序列；** 首先对UMI进行聚类，校正UMI的自身错误，在2对每个校正过的UMI cluster的序列进行比对，构建一条一致性序列，称为一个clone；

**3 Clonal Assignment** 对clone进行比对分析，相同的clone归类为一个clonetype，并统计每个clonetype含有的clone数目；

**4 注释** 采用IMGT数据库对鉴定的clonetype进行注释，鉴定出VDJ组合和CDR3序列等；

**5 V(D)J分析** 对不同V(D)J的使用频率和相互组合进行统计分析，并分析相关CDR3的长度；

**6 相似性和多样性评估** 对样本的Rankabundance变化趋势进行分析，并采用Hill指数的方式去评估样本的多样性；

**7 样本间的相关性分析** 对于多样本分析的情况，采用多方法多角度进行比较分析。

注：此分析流程图包括该产品的所有分析内容，本项目具体分析内容以此报告为准。

## 2 分析结果

### 2.1 数据概况

使用内部撰写的程序对原始的测序数据进行质量控制，获得Clean Data，并通过UMI对数据进行校正，校正后的数据通过数据库比对进行clone鉴定，对鉴定的clone做以下处理：

- (A)氨基酸个数大于等于4；
- (B)核酸长度是3的倍数；
- (C)核酸序列中不包含中止密码子。

详细的数据统计信息如下表所示：

表 2-1 数据统计信息

ID	Sample	ReadPair	FilteredReads	ReadUtilization (%)	Clones	Clonetypes	UniqueV	UniqueJ	UniqueVDJ	UniqueCDR3aa	UniqueCDR3nt
1	WAS9CD8TCM	3526218	2997821	85.02	4467	3446	22	12	385	3446	3444
2	WAS6CD8TCM	4424167	3305488	74.71	2236	801	23	12	265	801	799
3	WAS4CD8TCM	3805857	3012573	79.16	3047	1623	23	12	329	1623	1622
4	WAS8CD8TCM	3640802	3211601	88.21	4357	3337	23	12	377	3337	3332
5	WAS4CD4TEM	3091268	2446609	79.15	2751	1245	21	13	309	1245	1244
6	WAS6CD4TEM	3991945	3474606	87.04	3756	1201	22	13	307	1201	1192
7	WAS8CD4TEM	6429432	5595099	87.02	1438	616	23	13	242	616	610
8	WAS9CD4TEM	4594889	3913616	85.17	3348	1492	22	13	338	1492	1486

说明 Clones：有效克隆数；Clonetypes：clone的种类数；Unique V：不同 V 基因使用种类数；Unique J：不同 J 基因使用种类数；Unique VDJ：独特的 V/D/J 基因组合数；Unique CDR3aa：独特的 CDR3 氨基酸序列数；Unique CDR3nt：独特的 CDR3 核酸序列数。

随机挑选10000条原始reads做测序质量分布图如下图所示：

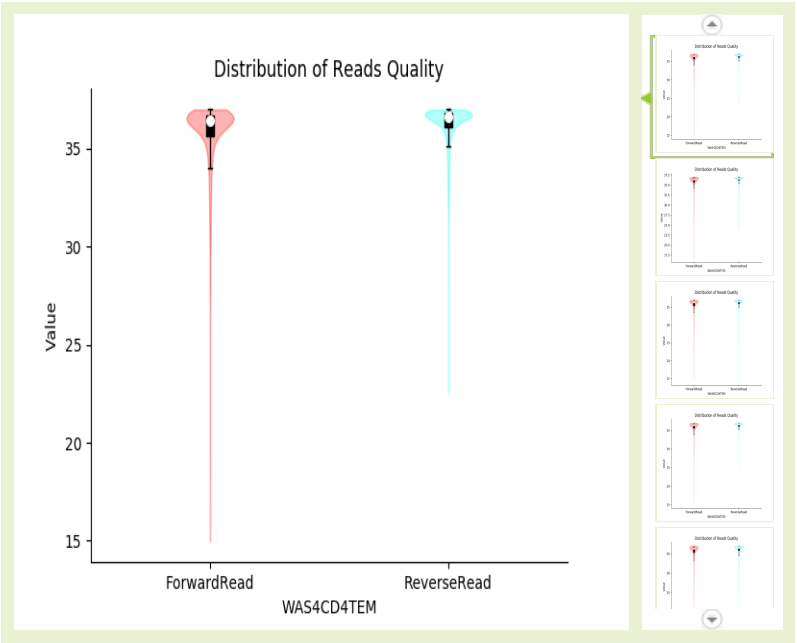


图 2-1 样品测序质量统计图

说明 横坐标为Reads类型，纵坐标为测序质量值。上图属于箱线图的一种变体，其与核密度图相结合。白点为中位数，黑色盒型的范围是下四分位点到上四分位点，细黑线表示须。外部形状即为核密度估计。一般情况下，中位数value值大于30，且密度分布集中在中位数值以上，即表明数据质量较高。

2.2 克隆的鉴定与计数

将样本测序得到的序列与IMGT数据库中记录的VDJ基因参考序列进行比对。根据每条序列的比对结果，确定相应的VDJ基因的使用情况和具体的CDR3区域序列，并过滤掉不符合条件或者没有功能的受体序列，统计结果部分如表 2-2

表 2-2 克隆的鉴定与计数

clonetype	cloneCount	cloneFraction	bestVGene	bestDGene	bestJGene	bestVFamily	bestDFamily	bestJFamily	VJCombination	VDJCombinat
2	5	0.00347705146036	TRBV20	TRBD1	TRBJ2-7	TRBV20	TRBD1	TRBJ2	TRBV20,TRBJ2-7	TRBV20,TRBD1,T7
3	25	0.0173852573018	TRBV5	TRBD1	TRBJ1-1	TRBV5	TRBD1	TRBJ1	TRBV5,TRBJ1-1	TRBV5,TRBD1,TF
4	3	0.00208623087622	TRBV12-1	TRBD1	TRBJ1-3	TRBV12	TRBD1	TRBJ1	TRBV12-1,TRBJ1-3	TRBV12-1,TRBD1,TRB.
6	28	0.019471488178	TRBV13-2	TRBD2	TRBJ2-4	TRBV13	TRBD2	TRBJ2	TRBV13-2,TRBJ2-4	TRBV13-2,TRBD2,TRB.
7	1	0.000695410292072	TRBV3	TRBD2	TRBJ2-4	TRBV3	TRBD2	TRBJ2	TRBV3,TRBJ2-4	TRBV3,TRBD2,TF
8	10	0.00695410292072	TRBV1	TRBD1	TRBJ2-5	TRBV1	TRBD1	TRBJ2	TRBV1,TRBJ2-5	TRBV1,TRBD1,TF
9	12	0.00834492350487	TRBV13-3	TRBD1	TRBJ1-2	TRBV13	TRBD1	TRBJ1	TRBV13-3,TRBJ1-2	TRBV13-3,TRBD1,TRB.
10	1	0.000695410292072	TRBV13-2	TRBD1	TRBJ1-4	TRBV13	TRBD1	TRBJ1	TRBV13-2,TRBJ1-4	TRBV13-2,TRBD1,TRB.
11	2	0.00139082058414	TRBV13-3	TRBD1	TRBJ1-1	TRBV13	TRBD1	TRBJ1	TRBV13-3,TRBJ1-1	TRBV13-3,TRBD1,TRB.
12	19	0.0132127955494	TRBV1	TRBD2	TRBJ2-5	TRBV1	TRBD2	TRBJ2	TRBV1,TRBJ2-5	TRBV1,TRBD2,TF
13	1	0.000695410292072	TRBV15	TRBD2	TRBJ1-1	TRBV15	TRBD2	TRBJ1	TRBV15,TRBJ1-1	TRBV15,TRBD2,T1
14	5	0.00347705146036	TRBV29	TRBD2	TRBJ2-5	TRBV29	TRBD2	TRBJ2	TRBV29,TRBJ2-5	TRBV29,TRBD2,T5
15	1	0.000695410292072	TRBV13-3	TRBD1	TRBJ2-4	TRBV13	TRBD1	TRBJ2	TRBV13-3,TRBJ2-4	TRBV13-3,TRBD1,TRB.
16	3	0.00208623087622	TRBV31	TRBD1	TRBJ1-1	TRBV31	TRBD1	TRBJ1	TRBV31,TRBJ1-1	TRBV31,TRBD1,T1

说明cloneCourt：某一clonetype的数量；cloneFraction：某一clonetype占整体的比值；bestV/D/JGene：最优的V/D/Jgene比对；bestV/D/JFamily：最优的V/D/J gene family；nSeqCDR3：CDR3区域的核苷酸序列；aaSeqCDR3：CDR3区域的氨基酸序列。

2.3 不同V/D/J基因使用频率

V、D、J的多样性构成了免疫组库的多样性，不同个体对V、D、J的使用频率是存在差异的，研究使用频率的差异对揭示样品免疫组库的差异具有重要意义，这里统计分析样本的V、D、J的使用频率，并绘制饼图，在legend区域展示前六个使用最为频繁的基因，如下图所示。

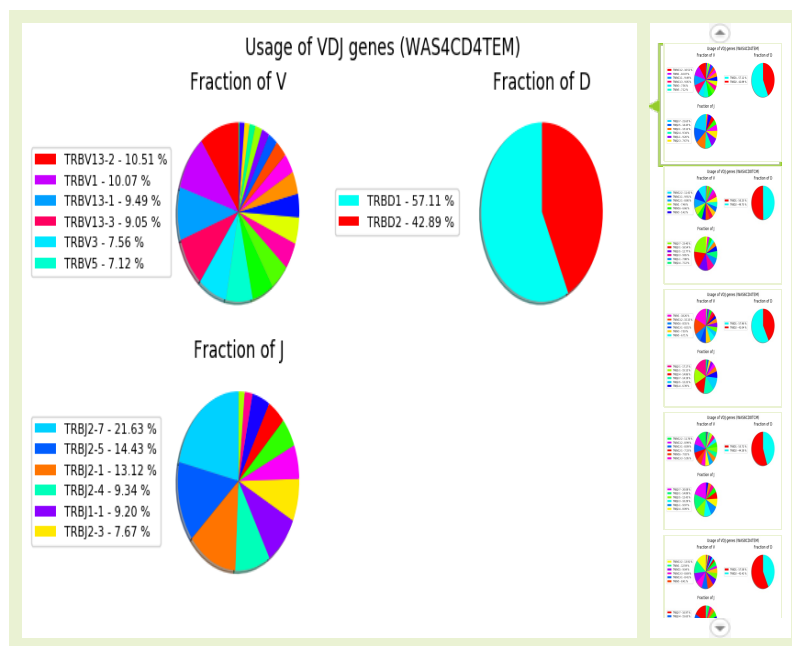


图 2-2 不同V/D/J基因使用频率统计图

## 2.4 不同V/D/J基因组合频率

不同V、D、J基因的组合极大的丰富了免疫组库的多样性，其中以VJ和VDJ的组合尤为重要，这里统计样本的组合方式和频率，绘制VJ组合的3D图（展示前30个最高频次Vgene），如下图：

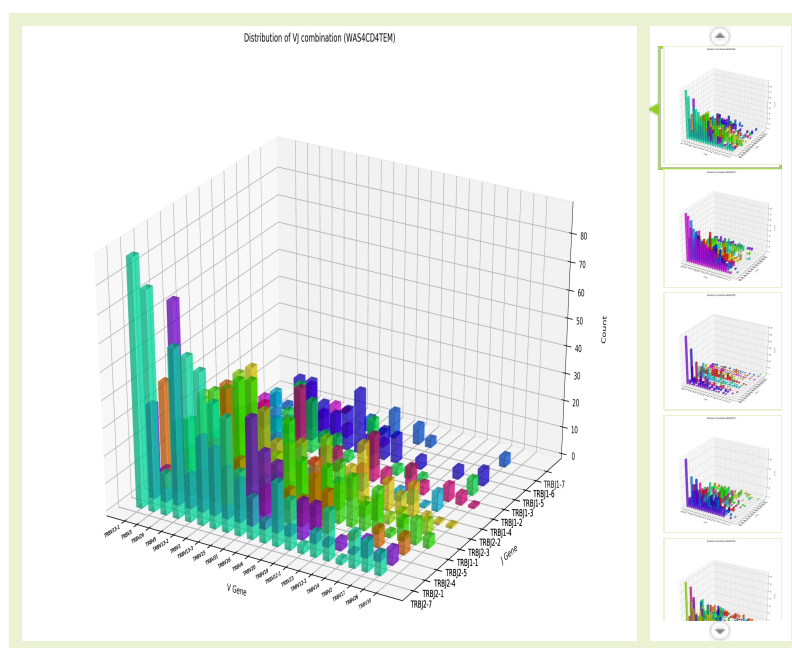


图 2-3 VJ Combination统计图

说明 X轴为V gene，Y轴为J gene，Z轴为相应组合的计数。

## 2.5 CDR3的长度分布

统计每个样本中Clone所包含的CDR3区域核酸和氨基酸长度分布信息，并绘制柱状图，充分展示样本的特征以及样本间的差异。

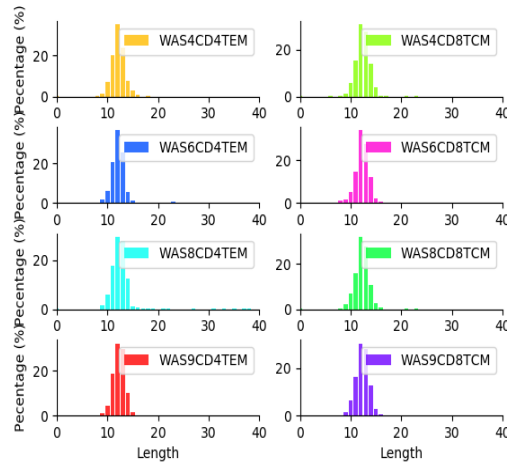


图2-4 样本的CDR3氨基酸长度分布图 ([点击大图](#))

说明 横坐标为长度，纵坐标为频率。

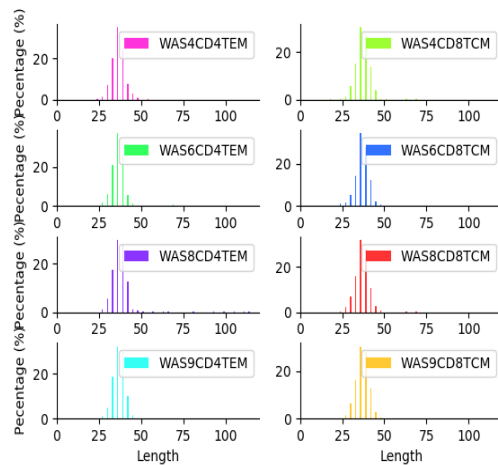


图2-5 样本的CDR3核苷酸长度分布图 ([点击大图](#))

说明 横坐标为长度，纵坐标为频率。

## 2.6 丰富度和多样性评估

评估免疫组库的丰富度和多样性是免疫组分析重要的工作，也最能体现出不同样本之间的差异。这里分析了样本的Rank abundance变化趋势以反映不同样本的丰富度关系，同时利用Hill指数，采用抽样的方法对样本的多样性特征进行评估。

Rank-Abundance 曲线是分析多样性的一种方式。构建方法是统计不同样本中，每一个CDR3区所含的序列数，将CDR3区按丰度（CDR3区所含有的序列条数）由大到小等级排序，再以CDR3区丰度高低的等级为横坐标，以每个CDR3区所含的序列丰度的频率为纵坐标做图。Rank-Abundance 曲线可用来解释多样性的两个方面：即CDR3区的丰度和不同CDR3区序列类型的均匀度。在水平方向，CDR3区的丰度由曲线的宽度来反映，CDR3区的丰度越高，曲线在横轴上范围越大；曲线的形状（平滑程度）反映了样本中CDR3区序列类型的均匀度，曲线越陡峭，样本中CDR3区序列类型的分布越不均匀。反之则趋于均匀。

Hill指数是连续变量 $q$ 的函数，能够直接反映三个样本多样性指数，也就是丰富度（Richness）、香浓指数（Shannon index）和Simpson index。当 $q=0$ 时，Hill指数也就是Richness值；当 $q=1$ 时，Hill值是Shannon index的指数值；当 $q=2$ 时，Hill值是Simpson index的倒数；当 $q$ 趋近于无穷大时，Hill指数能够反映出样本中最大的成分频率。



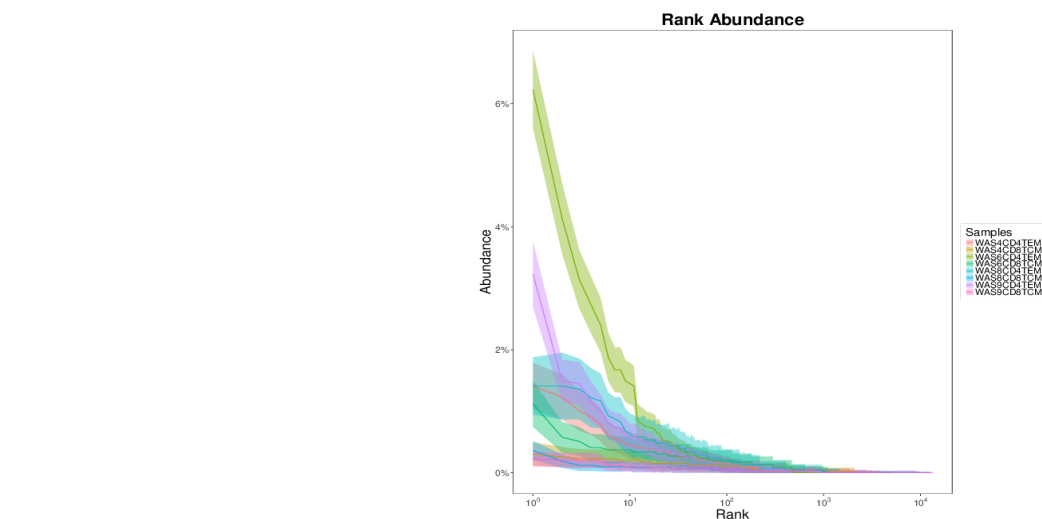


图2-6 样本Rank abundance分布曲线图 ([点击大图](#))

说明 横坐标表示CDR3丰度编号，纵坐标表示该编号对应的CDR3丰度频率，阴影区表示采用Bootstrap方法的0.95置信区间。

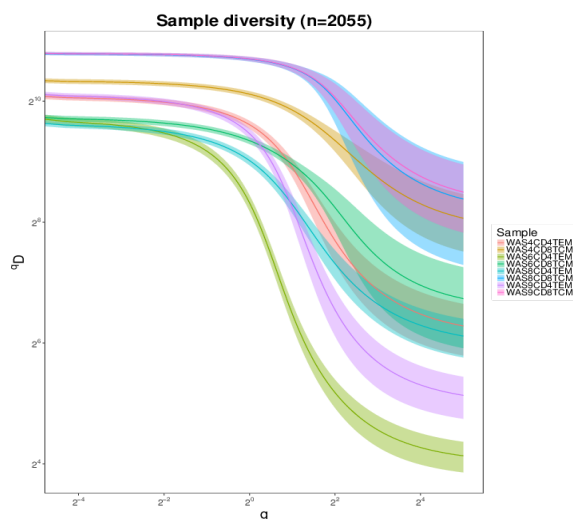


图2-7 样本Diversity分布曲线图 ([点击大图](#))

说明 横坐标表示不同q值，纵坐标表示该q值下样本的多样性。

丰富度 (Richness)、香浓指数 (Shannon index) 和Simpson index 这三个指数都是描述样本多样性的指标。从最初的针对生态学中生态群落多样性的使用，到后来应用到动物体内的微生物生态多样性、个体CDR3 区序列多样性评估等研究领域。

Richness：种的数目或丰富度 (species richness) 最初用于描述一个群落或生境中物种数目的多寡。Poole (1974) 认为只有这个指标才是唯一真正客观的多样性指标.在统计种的数目的时候,需要说明多大的面积,以便比较.在多层次的森林群落中必须说明层次和径级,否则是无法比较的。

Simpson：用来估算样本中微生物多样性指数之一，由Edward Hugh Simpson (1949) 提出，在生态学中常用来定量描述一个区域的生物多样性。Simpson 指数值越大，说明群落多样性越低。

Shannon：用来估算样本中微生物多样性指数之一。它与Simpson 多样性指数常用于反映alpha 多样性指数。Shannon 值越大，说明群落多样性越高。

样本 Diversity 分布曲线图就是把以上三种主流的、科研上使用得较多的多样性指数整合在一起进行分析。数值越高，则多样性越好。也可以理解为，同时比较横轴 $q=0$ ， $q=1$ ， $q=2$  时各个样本的纵坐标数值，纵坐标数值越高多样性越高。而由于Simpson index 越高多样性越低，所以这里分析的是其倒数。

## 2.7 样本间相关性分析

对于两个样本之间的比较，可以从多个角度评估两者件的相关性和相似性，这里采用评估样本间CDR3序列的相似性和VDJ基因以及相互组合的共

线性去评估样本之间的相关性。我们可以通过线性回归的方式考察组间差异及组内重复情况。样品间相关系数一般用R2表示，是检验实验可靠性和样本选择合理性的重要指标，R²最大值为1。R²的值越接近1，说明回归直线对观测值的拟合程度越好；反之，R²的值越小，说明回归直线对观测值的拟合程度越差。相关系数越接近1表明样品之间的相似度越高，反之亦然。人类组织来源样品无法获得严格意义上的生物学重复，通过增加样本量可以提高分析结果的统计意义。

2.7.1 样本间CDR3序列的相似性

比较样本间CDR3的核苷酸和氨基酸序列的重叠情况以评估样本间CDR3序列的相似性，根据 CDR3 的核苷酸序列或氨基酸序列的重叠情况所绘制的韦恩图，是基于特定的CDR3区其Clonotype 的氨基酸序列或核苷酸序列，在样本间的数目情况差异所绘制的。样本间存在相同的Clonotype 氨基酸序列，则在韦恩图的重叠区显示，韦恩图如下（比较样本数多于6个是不绘制韦恩图）：

表 2-3 CDR3的氨基酸序列的重叠情况

Clonotype	WAS9CD8TCM	WAS6CD8TCM	WAS4CD8TCM	WAS8CD8TCM	WAS4CD4TEM	WAS6CD4TEM	WAS8CD4TEM	WAS9CD4
ASSLAGIYAEQF	0.0199927299164	0	0	0	0	0	0	0
ASSDGDWYEQY	0.0130861504907	0	0	0	0	0	0	0
ASSPTGGGQDTQY	0.00799709196656	0	0	0	0	0	0	0
TCSGDWGGQNTLY	0.00763358778626	0	0	0	0	0	0	0
TCSAGLGVGDTQY	0.00690657942566	0	0	0	0	0	0	0
GARTGTNTEVF	0.00654307524537	0	0	0	0	0	0	0
SSSQSSGGAGQDTQY	0.00617957106507	0	0	0	0	0	0	0
ASSQAGTEDTQY	0.00581606688477	0	0	0	0	0	0	0
AWRGTGNYAEQF	0.00545256270447	0	0	0	0	0	0	0
ASSLGPAETLY	0.00545256270447	0	0	0	0	0	0	0
TCSAQDSYEQY	0.00545256270447	0	0	0	0	0	0	0
ASSGPGTISNERLF	0.00545256270447	0	0	0	0	0	0	0
ASSPGTFQDTQY	0.00508905852417	0	0	0	0	0	0	0
ASSQGDQNTLY	0.00508905852417	0	0	0	0	0	0	0

2.7.2 样本间VDJ基因以及相互组合的共线性

该分析均基于样本间 V 基因，VJ 基因组合，VDJ 基因组合的频率差异表格来绘制的相关性图，图中的横坐标和纵坐标均为频率值。若两个样本的V 基因,VJ基因组合, VDJ 基因组合其各自的频率相差较大，则会在图上标记出来。

统计样本间V基因的频率差异，给出具体的差异情况，并根据频率绘制共线性图，计算相关共线性，结果如下所示：

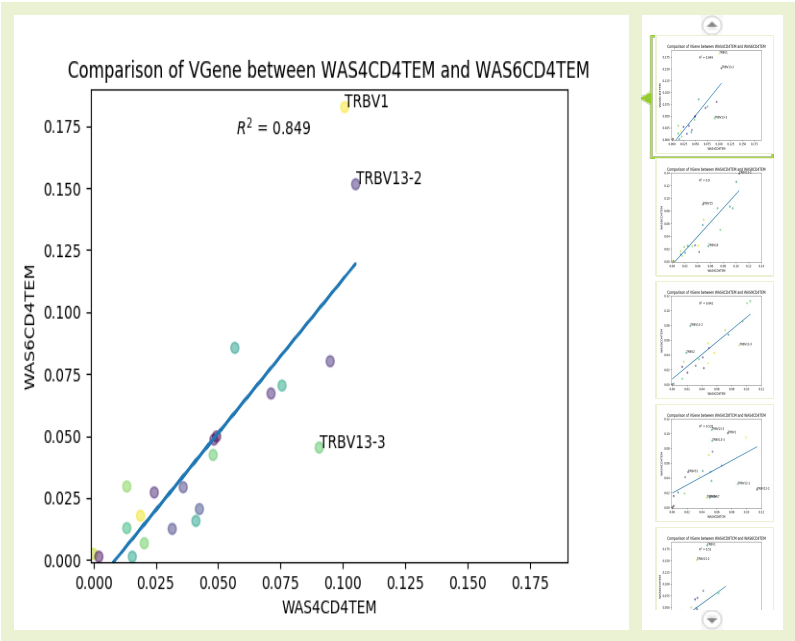


图 2-10 样本间V基因共线性

表 2-4 样本间V基因的频率差异

VGene	WAS9CD8TCM	WAS6CD8TCM	WAS4CD8TCM	WAS8CD8TCM	WAS4CD4TEM	WAS6CD4TEM	WAS8CD4TEM	WAS9CD4TEM
TRBV13-2	0.1237	0.1141	0.1234	0.1394	0.0616	0.0899	0.0884	0.0685
TRBV13-1	0.0956	0.0799	0.0763	0.0791	0.0893	0.0749	0.0908	0.0868
TRBV13-3	0.0908	0.055	0.0844	0.0603	0.0555	0.0712	0.0653	0.0667
TRBV5	0.0787	0.0699	0.0698	0.0804	0.0567	0.02	0.0381	0.0421
TRBV1	0.0779	0.1091	0.0877	0.0858	0.0647	0.04	0.0369	0.0537
TRBV3	0.0755	0.0716	0.0617	0.0777	0.0548	0.0674	0.0455	0.0566
TRBV19	0.0594	0.0608	0.0763	0.061	0.0444	0.0487	0.039	0.0415
TRBV31	0.0562	0.0666	0.0633	0.0402	0.0191	0.0287	0.0168	0.0203
TRBV15	0.0538	0.0774	0.1006	0.0657	0.0487	0.0487	0.0494	0.0508
TRBV16	0.049	0.0716	0.0276	0.0416	0.0684	0.0674	0.0908	0.0696
TRBV4	0.045	0.0425	0.0276	0.0369	0.0474	0.0512	0.0396	0.0409
TRBV20	0.0305	0.0275	0.0276	0.0302	0.0173	0.0062	0.0123	0.0163
TRBV29	0.0289	0.0266	0.0195	0.0315	0.0345	0.0562	0.0593	0.0427
TRBV12-1	0.0281	0.015	0.0325	0.0228	0.0912	0.0749	0.107	0.1042

统计样本间VJ基因组合的频率差异，给出具体的差异情况，并根据频率绘制共线性图，计算相关共线性，结果如下所示：

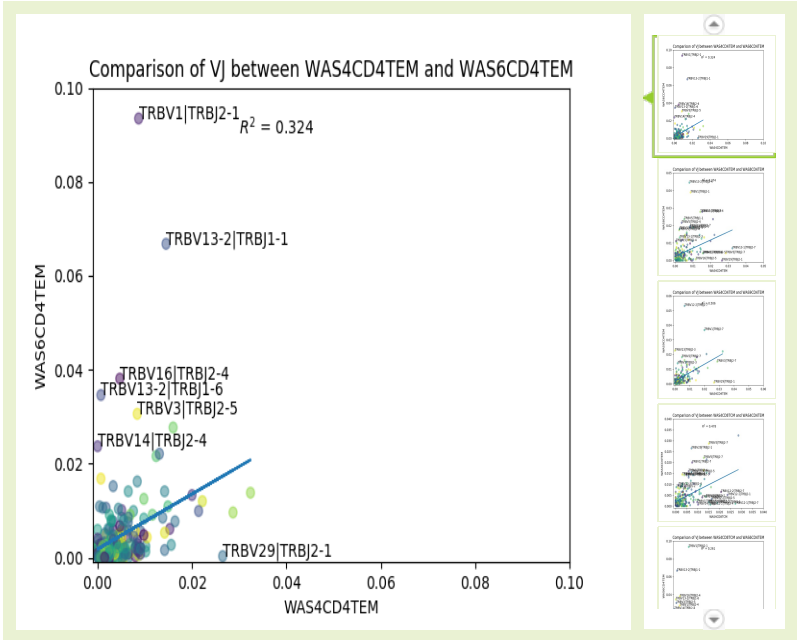


图 2-11 样本间VJ基因共线性

表 2-5 样本间VJ基因组合的频率差异

VJGene	WAS9CD8TCM	WAS6CD8TCM	WAS4CD8TCM	WAS8CD8TCM	WAS4CD4TEM	WAS6CD4TEM	WAS8CD4TEM	WAS9CD4TEM
TRBV3 TRBJ2-7	0.0249	0.0167	0.0065	0.0168	0.016	0.0262	0.012	0.0151
TRBV13-2 TRBJ2-7	0.0249	0.0175	0.0211	0.0282	0.0179	0.0125	0.0195	0.0157
TRBV13-1 TRBJ2-7	0.0249	0.0142	0.0081	0.0181	0.024	0.0112	0.0216	0.0165
TRBV5 TRBJ2-7	0.0225	0.0092	0.0179	0.0147	0.0129	0.0075	0.0108	0.0087
TRBV13-2 TRBJ2-5	0.0201	0.0183	0.0195	0.0174	0.0074	0.0175	0.0129	0.0107
TRBV13-3 TRBJ2-7	0.0185	0.0083	0.013	0.0101	0.0142	0.015	0.0132	0.0116
TRBV15 TRBJ2-7	0.0161	0.02	0.0146	0.0188	0.0148	0.005	0.0153	0.0174

TRBV13-2 TRBJ1-1	0.0145	0.0192	0.0049	0.0141	0.0031	0.0087	0.0069	0.0044
TRBV1 TRBJ2-7	0.0137	0.0158	0.0146	0.0201	0.0068	0.0075	0.0069	0.0102
TRBV13-1 TRBJ2-4	0.0129	0.0133	0.0146	0.0127	0.008	0.0062	0.0087	0.011
TRBV13-1 TRBJ2-1	0.0129	0.0067	0.0049	0.008	0.0129	0.0125	0.0183	0.0122
TRBV31 TRBJ2-1	0.012	0.0058	0.0032	0.0047	0.0037	0.0062	0.0018	0.0035
TRBV13-2 TRBJ2-3	0.012	0.0042	0.0114	0.0107	0.0043	0.0075	0.0111	0.0075
TRBV31 TRBJ2-7	0.0112	0.0142	0.0195	0.006	0.0043	0.005	0.0027	0.0046

统计样本间VDJ基因组合的频率差异，并根据频率绘制共线性图，计算相关共线性，结果如下所示：

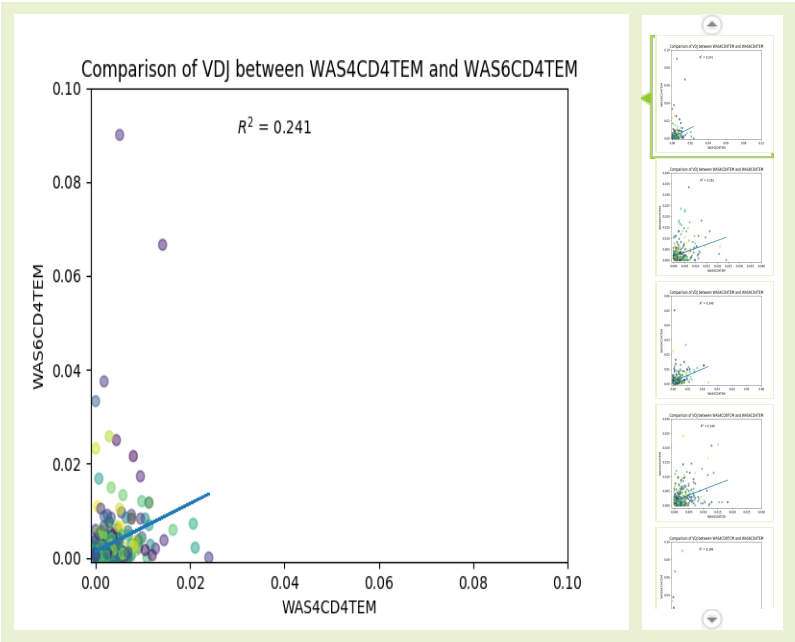


图 2-12 样本间VDJ基因共线性

表 2-6 样本间VDJ基因组合的频率差异

VDJGene	WAS9CD8TCM	WAS6CD8TCM	WAS4CD8TCM	WAS8CD8TCM	WAS4CD4TEM	WAS6CD4TEM	WAS8CD4TEM	WAS9CD4TE
TRBV3 TRBD2 TRBJ2-7	0.0185	0.0092	0.0049	0.0127	0.0129	0.02	0.0081	0.0107
TRBV13-2 TRBD2 TRBJ2-7	0.0177	0.01	0.0179	0.0154	0.0148	0.01	0.0123	0.011
TRBV13-2 TRBD1 TRBJ1-1	0.0137	0.0183	0.0049	0.0127	0.0031	0.0075	0.0054	0.0044
TRBV5 TRBD2 TRBJ2-7	0.0129	0.0017	0.013	0.0101	0.0062	0.0037	0.0075	0.0052
TRBV13-2 TRBD2 TRBJ2-5	0.0129	0.0058	0.0146	0.0094	0.0049	0.01	0.0075	0.009
TRBV13-1 TRBD1 TRBJ2-7	0.0129	0.01	0.0016	0.0067	0.0092	0.0025	0.0069	0.0064
TRBV13-3 TRBD2 TRBJ2-7	0.012	0.0058	0.0097	0.0054	0.0086	0.0112	0.0066	0.0032
TRBV13-1 TRBD2 TRBJ2-7	0.012	0.0042	0.0065	0.0114	0.0148	0.0087	0.0147	0.0102
TRBV3 TRBD1 TRBJ1-1	0.0104	0.0058	0.0049	0.008	0.0018	0.005	0.0021	0.0035
TRBV15 TRBD2 TRBJ2-7	0.0104	0.0125	0.0065	0.0107	0.0123	0.0037	0.009	0.0122
TRBV5 TRBD1 TRBJ2-7	0.0096	0.0075	0.0049	0.0047	0.0068	0.0037	0.0033	0.0035
TRBV13-2 TRBD2 TRBJ2-1	0.0096	0.0083	0.0081	0.0094	0.008	0.0137	0.0111	0.0104

TRBV13-2 TRBD2 TRBJ2-3	0.0096	0.0042	0.0097	0.0054	0.0031	0.005	0.0078	0.0061
TRBV4 TRBD2 TRBJ2-1	0.0088	0.0025	0	0.0027	0.0043	0.0012	0.0039	0.0046

从基因家族的角度可以在更加宏观的角度评估样本间的相似性，这里针对V gene绘制共线性图，如下图所示：

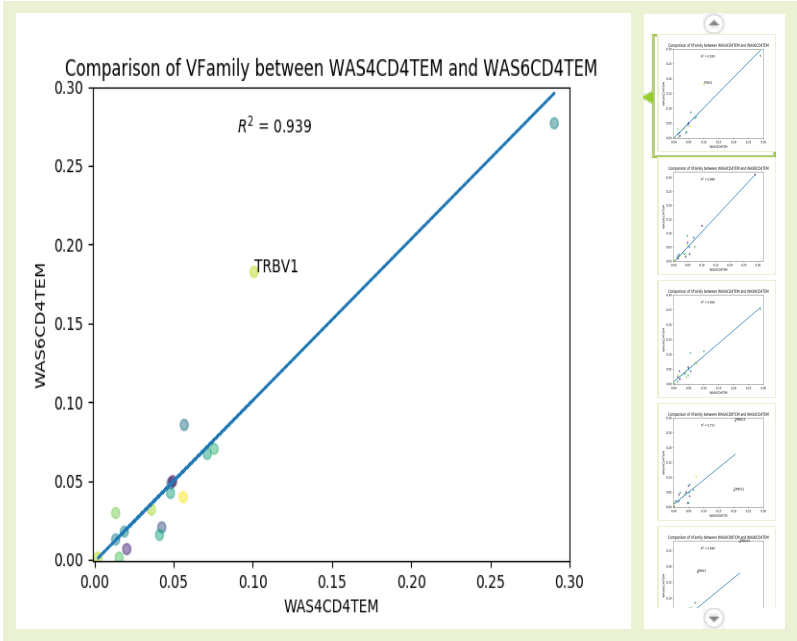


图 2-13 样本间VFamily基因共线性

2.7.3 样本间差异假设检验分析

为探究样本间的差异性，采用双样本T检验的方法对样本的V基因频率、VJ组合频率、VDJ组合频率和Clonotype频率进行假设检验分析，部分结果如下表所示：

表 2-7 样本间假设检验分析

Clonotype	WAS9CD8TCM	WAS6CD8TCM	WAS4CD8TCM	WAS8CD8TCM	WAS4CD4TEM	WAS6CD4TEM	WAS8CD
ASSRDSYEQY	0.00327153762268	0.00292864749734	0.00208623087622	0	0	0	0.00045903
ASSFTGGDTEVF	0.00254452926209	0.00452609158679	0	0.00567502986858	0	0	0
ASSRGTGGYEYQY	0	0	0	0	0	0.000447227191413	0.00045903
ASSPGQTNSDYT	0	0	0	0	0.000328191663932	0	0.00022951
ASSLSGLGGAEQF	0	0	0	0	0.000328191663932	0	0.00022951
ASGDWQGDTQY	0	0	0	0	0	0.000447227191413	0.00022951
ASSPGHYEQY	0	0	0	0	0.000328191663932	0	0.00045903
ASSLGWDYEYQY	0	0	0	0	0.000328191663932	0	0.00022951
ASSSDWGSTGQLY	0	0	0	0	0.000328191663932	0	0.00045903
ASTGRAETLY	0	0	0	0	0.000328191663932	0	0.00068854
ASSQDWQGDTQY	0	0	0	0	0.000328191663932	0	0.00022951
ASSPGTGGAEEQF	0	0	0	0	0.000328191663932	0	0.00045903
ASSPDWQGDTQY	0	0	0	0	0.000328191663932	0	0.00045903
ASSLDSSGNTLY	0	0	0	0	0.000656383327863	0.000894454382826	0.00022951

2.7.4 样本间PCA分析

PCA 分析(Principal Component Analysis)，即主成分分析，是一种对数据进行降维从而简化数据的分析技术，这种方法可以有效的找出数据中最“主要”的元素和结构，去除噪音和冗余，将原有的复杂数据降维，揭示隐藏在复杂数据背后的简单结构。其优势在于简单且无参数限制。

通过分析不同样本间的基因及其组合频率和Clonotype频率可以反映样本间的差异和距离。PCA 运用方差分解，将多组数据的差异反映在二维坐标图上，坐标轴取能够最大反映方差值的两个特征值。如样本组成越相似，反映在PCA 图中的距离越近。不同特性的样本可能表现出分散和聚集的分布情况，PCA 结果中对样本差异性解释度最高的两个或三个成分可以用于对假设因素进行验证。

对V基因频率、VJ组合频率、VDJ组合频率和Clonotype频率进行主成分计算，取第一、二位的主成分值进行作图，结果如下图所示：

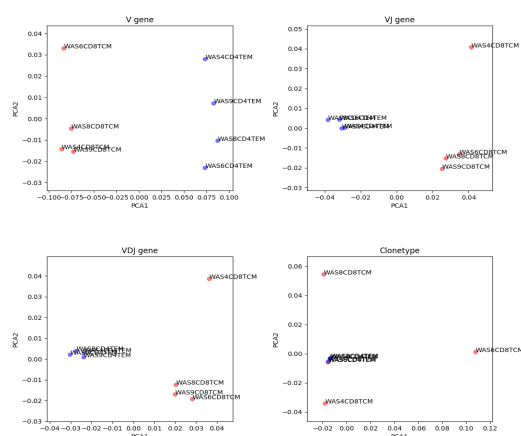


图 2-14 样本间PCA分析([点击大图](#))

### 3 参考文献

- [1] Bolotin, Dmitriy A., et al. "MiXCR: software for comprehensive adaptive immunity profiling." *Nature methods* 12.5 (2015): 380-381.
- [2] Gupta, Namita T., et al. "Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data." *Bioinformatics* 31.20 (2015): 3356-3358.
- [3] Vander Heiden, Jason A., et al. "pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires." *Bioinformatics* 30.13 (2014): 1930-1932.
- [4] Lefranc, Marie-Paule, et al. "IMGT®, the international ImMunoGeneTics information system®." *Nucleic acids research* 37.suppl\_1 (2008): D1006-D1012.
- [5] Ye, Jian, et al. "IgBLAST: an immunoglobulin variable domain sequence analysis tool." *Nucleic acids research* 41.W1 (2013): W34-W40.
- [6] Chunlin Wanga, Catherine M. Sandersb, Qunying Yang et al. High throughput sequencing reveals a complex pattern of dynamic interrelationships among human T cell subsets. *PANS*, 2010, vol. 107:1518–1523.
- [7] Jennifer Benichou, Rotem Ben-Hamo, Yoram Louzoun et al. 2Rep-Seq uncovering the immunological repertoire through next-generation sequencing. *Immunology*, 2011, 135: 183–191.
- [8] Jeroen W J van Heijst, Izaskun Ceberio, Lauren B Lipuma et al. Quantitative assessment of T cell repertoire recovery after hematopoietic stem cell transplantation. *Nature Medicine*, 2013,VOL.19: 372-379.
- [9] Miran Jang, Poh-Yin Yew, Kosei Hasegawa et al. Characterization of T cell repertoire of blood,tumor, and ascites in ovarian cancer patients using next generation sequencing. *Oncolmmunology*, 2015, Vol.4:e1030561-1-10.
- [10] Evaggelia Liaskou, Eva Kristine Klemsdal Henriksen, Kristian Holm et al. High-throughput T-cell receptor sequencing across chronic liver diseases reveals distinct disease-associated repertoires. *Hepatology*, 2015,1-10.
- [11] X-L Hou, LWWang, Y-L Ding, et al. Current status and recent advances of next generation sequencing techniques in immunological repertoire. *Genes and Immunity*, 2016, 1–12. [12] Bates S T, Clemente J C, Flores G E, et al. Global biogeography of highly diverse protistan communities in soil[J]. *The ISME journal*. 2013, 7(3): 652-659.

## 4 备注

### 4.1 结果文件说明

- 1、Summary.txt：项目样本的分析概况。
- 2、\*.CloneFilter.txt：样本的clone序列以及比对VDJ的详细信息。
- 3、\*\_VDJUsage.xls：样本的V、D、J基因的使用频率统计信息。
- 4、\*\_VDJCombination.xls：样本的VJ Combination和VDJ Combination频率统计信息。
- 5、VFraction.txt：样本间V基因的频率比较。
- 6、VJFraction.txt：样本间VJ基因组合的频率比较。
- 7、VDJFraction.txt：样本间VDJ基因组合的频率比较。
- 8、ClonotypeFraction.txt：样本间clonotype的频率比较。
- 9、MixDiffAnalysis.xls：样本间假设检验结果文件。

结果文件建议使用Excel或者EditPlus等专业文本编辑器打开。

### 4.2 分析软件说明

Software	Version
Igblast	2.6.1
Mixcr	2.1.5
pRESTO	0.5.3
Alakazam	0.2.8
SHazaM	0.1.8

推荐使用火狐浏览器进行网页版结题报告浏览，下载地址：<http://www.firefox.com.cn/download/>