免疫组库结题报告

广州华银医学检验中心

2018年11月22日

Table of Contents

2 分析结果

3 参考文献

4 备注

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 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 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。

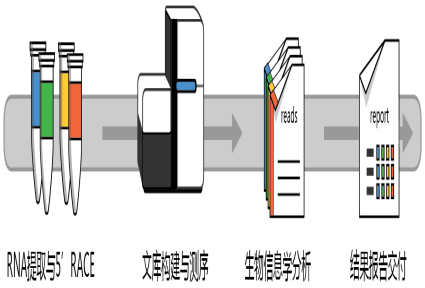

图 1-1 实验流程图

1.2 生物信息分析流程

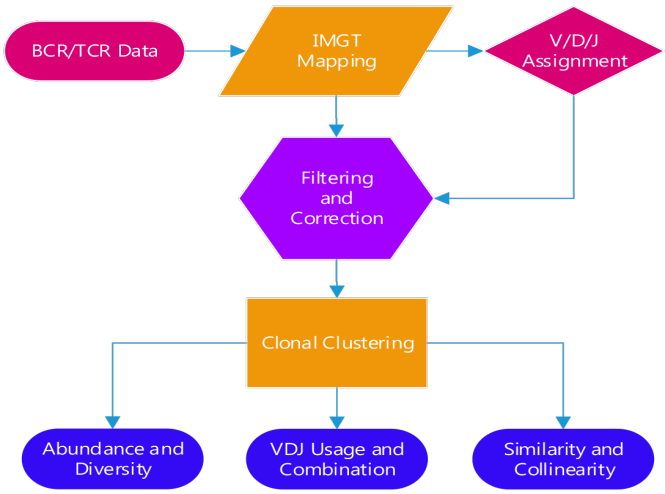

图 1-2 生物信息分析流程图

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 数据库比对 与IMGT数据库进行序列比对，过滤掉未比对上的Reads，保留得到比对结果的Reads，鉴定出相应的V(D)J基因；

3 Clonal Assignment 针对所有的受体序列鉴定出相应的克隆并计数；

4 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；

5 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；

6 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 测序数据质量评估

在进行数据过滤时，使用内部撰写的程序对原始的测序数据进行如下处理，获得Clean Data，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。
详细的质控统计信息如下表所示：

表 2-1 质控统计信息

ID	Sample	ReadPair	FilteredReads	Clone	ClonePercent(%)	UniqueV	UniqueJ	UniqueVDJ	UniqueCDR3aa	UniqueCDR3nt
1	Case3	5596590	5414981	4809818	85.94	94	6	7490	149265	159451
2	Case2	5567009	5379249	4779501	85.85	94	6	7860	182958	197879
3	Case1	8125693	7564635	6731389	82.84	95	6	7429	142673	148907
4	Control3	7468492	6908748	5899261	78.99	97	6	8668	224317	246118
5	Control1	4823812	4667194	4051375	83.99	95	6	7884	148892	161044
6	Control2	7556968	6911532	5983036	79.17	94	6	6821	81161	83888

说明 Clone (clean clone)：有效克隆数; Clone Percent (%)：有效克隆数占总Reads 的比例; 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：

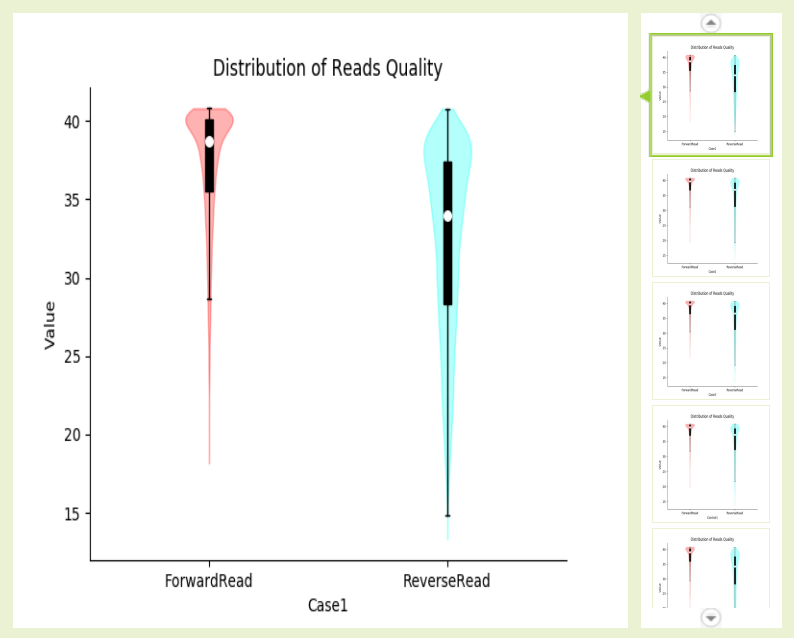

图 2-1 样品测序质量统计图

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

利用MIXCR软件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

cloneId	cloneCount	cloneFraction	bestVGene	bestDGene	bestJGene	bestVHitScore	bestDHitScore	bestJHitScore	bestVFamily	bestDFamily	bestJFamily	
0	13980	0.0033764967807349364	IGHV4-39	IGHD2-8	IGHJ4	307.6	35.0	145.7	IGHV4	IGHD2	IGHJ4	TGTGCGAGACA
1	7094	0.0017133668213543374	IGHV4-34	IGHD2-2	IGHJ4	317.1	45.0	130.0	IGHV4	IGHD2	IGHJ4	TGTGCGAGTGA
2	6769	0.0016348717245203708	IGHV3-73	IGHD4-11	IGHJ4	362.6	41.0	154.2	IGHV3	IGHD4	IGHJ4	TGTACTC
3	5900	0.0014249879117550876	IGHV4-4	IGHD3-10	IGHJ3	290.8	25.0	135.2	IGHV4	IGHD3	IGHJ3	TGTGC
4	5504	0.001329344655305085	IGHV4-34	IGHD2-2	IGHJ4	309.2	46.0	164.0	IGHV4	IGHD2	IGHJ4	TGTGCGGGTGA
5	5479	0.0013233065709332414	IGHV4-39	IGHD1-26	IGHJ4	242.6	30.0	164.3	IGHV4	IGHD1	IGHJ4	TGTGCGAGA
6	5206	0.0012573706895927094	IGHV4-39	IGHD3-3	IGHJ4	263.4	36.0	156.6	IGHV4	IGHD3	IGHJ4	TGTGCGAG
7	5120	0.0012365996793535674	IGHV3-74	IGHD6-6	IGHJ4	309.8	75.0	199.0	IGHV3	IGHD6	IGHJ4	TGTGCCCGAGAG
8	4860	0.0011738036018863942	IGHV5-51	IGHD3-16	IGHJ4	210.4	35.0	144.6	IGHV5	IGHD3	IGHJ4	TGTGCGCGGC
9	4844	0.0011699392278884144	IGHV4-39	IGHD4-23	IGHJ5	215.0	36.0	112.5	IGHV4	IGHD4	IGHJ5	TGTGCGAGACGGACGGTGC
10	4709	0.0011373335722804588	IGHV6-1	IGHD5-24	IGHJ4	152.7	30.0	160.3	IGHV6	IGHD5	IGHJ4	TGTGCTAGAGAGA
11	4268	0.001030821763961138	IGHV2-5	IGHD6-13	IGHJ4	361.0	62.0	198.9	IGHV2	IGHD6	IGHJ4	TGTGCACACACGAGAC
12	4226	0.0010206777822164407	IGHV4-34	IGHD7-27	IGHJ4	216.8	31.0	205.4	IGHV4	IGHD7	IGHJ4	TGTGCGAGAGGCTCTAT
13	3858	9.317971802629031E-4	IGHV3-64	IGHD4-11	IGHJ4	350.1	25.0	178.9	IGHV3	IGHD4	IGHJ4	TG

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/JHitScore：最优V/D/Jgene比对的得分；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legand区域展示前六个使用最为频繁的基因，如下图所示。

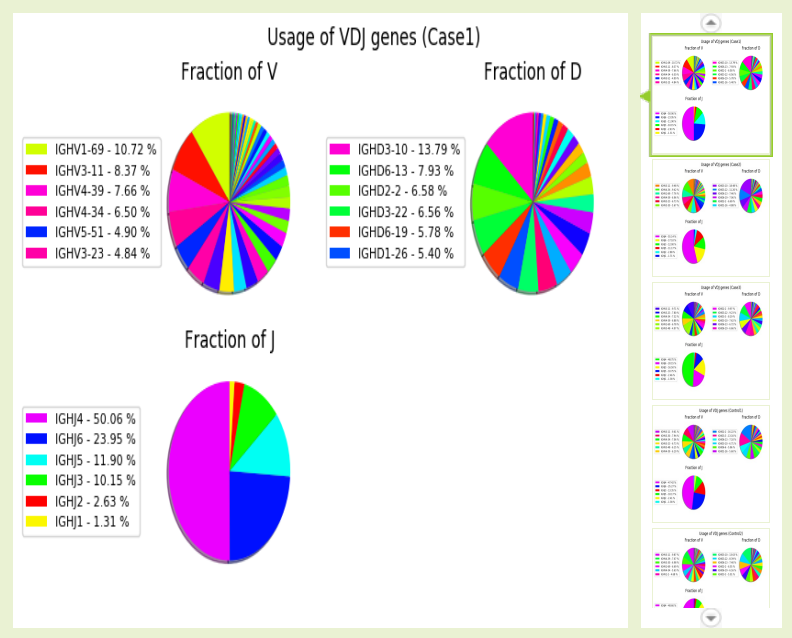

图 2-2 不同V/D/J基因使用频率统计图

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：

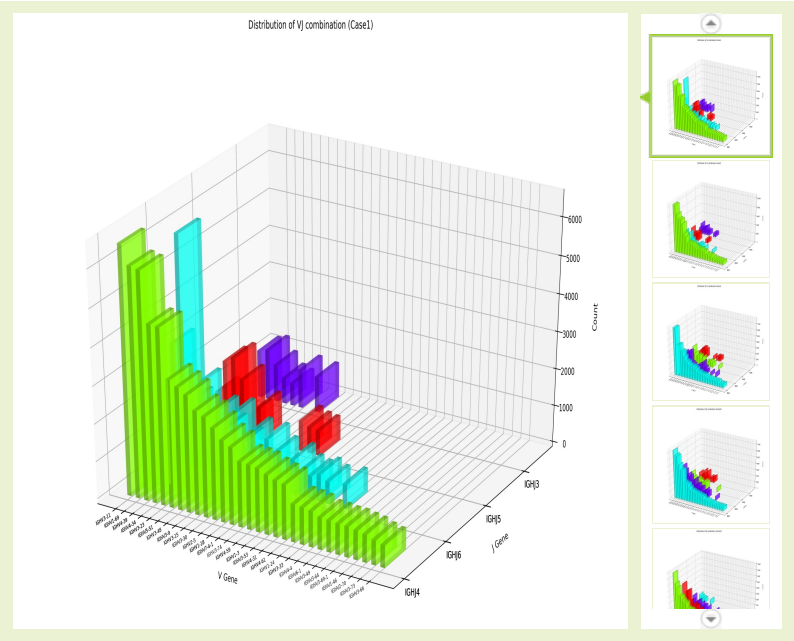

图 2-3 VJ Combination统计图

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，并绘制小提琴图，充分展示样本的特征以及样本间的差异。

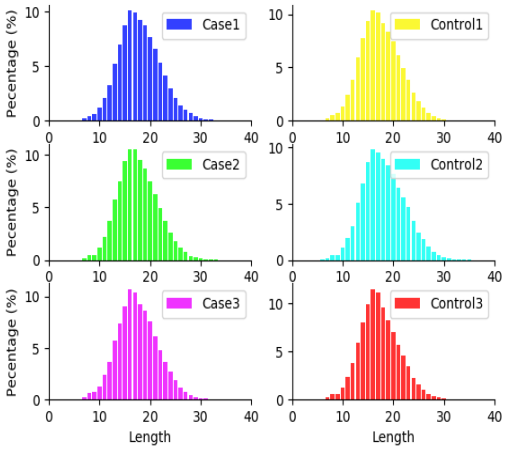

图2-4 样本的CDR3氨基酸长度分布图 (点击大图)

说明 横坐标为长度，纵坐标为频率。

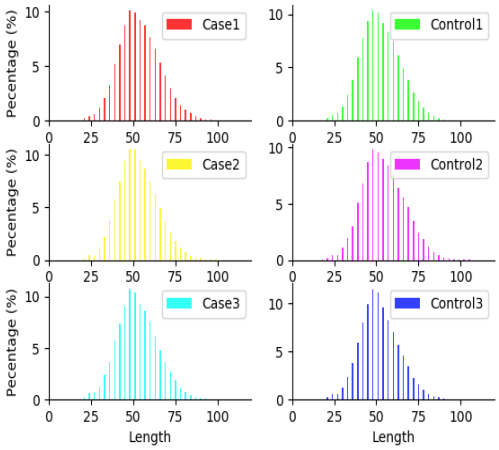

图2-5 样本的CDR3核苷酸长度分布图 (点击大图)

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。

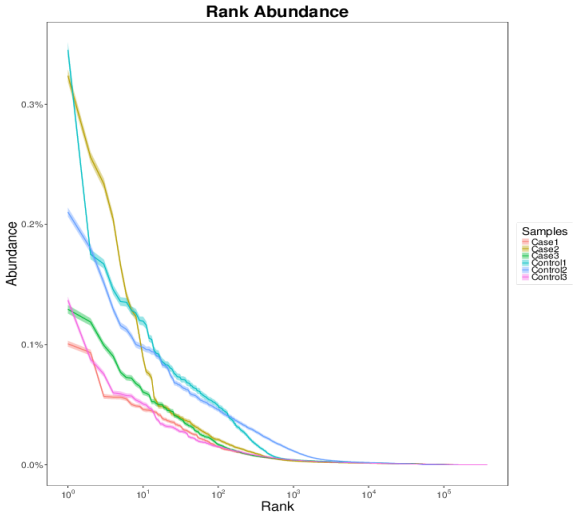

图2-6 样本Rank abundance分布曲线图 (点击大图)

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。

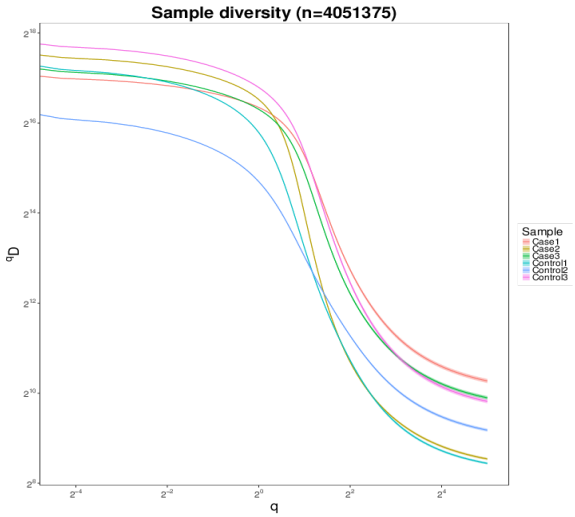

图2-7 样本Diversity分布曲线图 (点击大图)

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 HughSimpson (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，可以从多个角度比评估两者这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，R^2最大值为1。R^2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，R^2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Clonetype 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Clonetype 氨基酸序列，则在韦恩图的重叠区显示，韦恩图如下：

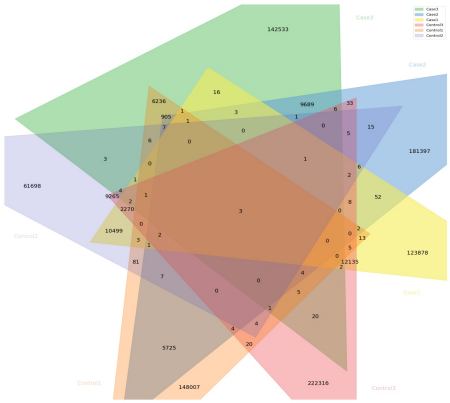

图 2-8 样本间CDR3核苷酸序列的相似性(点击大图)

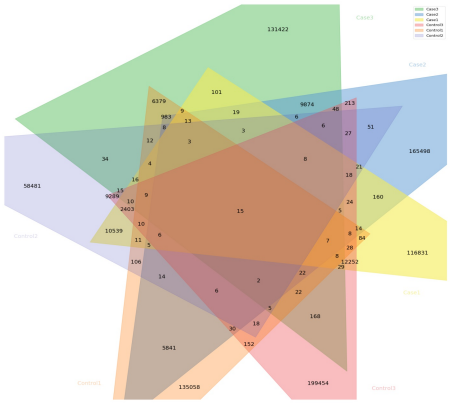

图 2-9 样本间CDR3氨基酸序列的相似性(点击大图)

统计CDR3的氨基酸序列的重叠情况以及相应频率，给出具体的clone差异，结果如下表所示：

表 2-3 CDR3的氨基酸序列的重叠情况

Clonetype	Case3	Case2	Case1	Control3	Control1	Control2
CAHRVGQTVFGLVINTDIFFDSW	0.000989143792929	0	0	0	2.80489682187e-06	1.65582652237e-06
CTTWNYGGKFGLQFVSHW	0.000917479398311	0	0	0	2.14492109908e-06	2.48373978355e-06
CARGDSRTEYYEYSGPEYFQQW	0.000559449337214	0	0	0	1.31995144559e-06	1.49024387013e-06
CARHREPYYGEFHDALNMW	0.000553319185332	0	0	0	1.64993930698e-06	1.49024387013e-06
CVRALNSNFDSW	0.000553173229335	0	0	0	1.64993930698e-07	1.8214091746e-06
CAHRRDGGLDRTYYYGMDVW	0.000538577629616	0	0	0	9.8996358419e-07	8.27913261183e-07
CAHKLERLYSFDYW	0.000494498918466	0	0	0	3.29987861397e-07	1.15907856566e-06
CAHRSSFFDYW	0.000480925010728	0	0	0	3.29987861397e-07	8.27913261183e-07
CARGGTRDGYNYVGYFDYW	0.000478297802778	0	0	0	3.29987861397e-07	2.15257447908e-06
CARHTPSYSFFRETAVSYDPDYW	0.000452609547274	0	0	0	6.59975722793e-07	6.62330608947e-07
CARLSGWSSGWYPDLW	0.000448376823355	0	0	0	3.29987861397e-07	1.65582652237e-07
CARDGPRTGYGDYVGGYYW	0.000446187483398	0	0	0	8.24969653491e-07	1.15907856566e-06

CARHFYYGSGSFYPSSKSHDYW	0.000440495199507	0	0	0	8.249696534 91e-07	1.655826522 37e-07
CAHRRSSSSGFDYW	0.000433927179634	0	0	0	9.8996358419e-07	4.9674795671e-07

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，图中的横坐标和众坐标均为频率值。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：

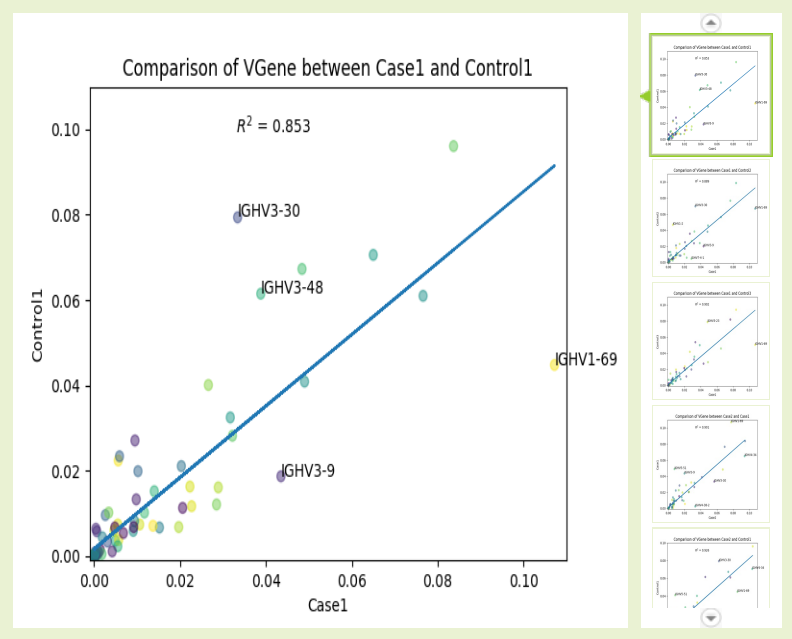

图 2-10 样本间V基因共线性

表 2-4 样本间V基因的频率差异

VGene	Case3	Case2	Case1	Control3	Control1	Control2
IGHV1-69	0.1072	0.0776	0.0678	0.0448	0.0669	0.0505
IGHV3-11	0.0837	0.0948	0.0971	0.0961	0.0987	0.0941
IGHV4-39	0.0766	0.0696	0.0688	0.061	0.0767	0.0819
IGHV4-34	0.065	0.0942	0.0732	0.0706	0.0563	0.0459
IGHV5-51	0.049	0.0076	0.0206	0.0409	0.0458	0.0289
IGHV3-23	0.0484	0.0673	0.078	0.0673	0.0382	0.0788
IGHV3-9	0.0435	0.0199	0.0355	0.0187	0.0201	0.0271
IGHV3-48	0.0388	0.0415	0.0497	0.0615	0.024	0.05
IGHV3-30	0.0334	0.0567	0.0391	0.0794	0.0698	0.0537
IGHV1-18	0.0322	0.0283	0.0335	0.0282	0.0384	0.0357
IGHV3-15	0.0317	0.0324	0.029	0.0325	0.0238	0.0323
IGHV2-5	0.0289	0.0219	0.0155	0.0161	0.0266	0.0252
IGHV7-4-1	0.0285	0.0138	0.0144	0.0121	0.0046	0.0198
IGHV4-59	0.0266	0.0321	0.036	0.0401	0.0357	0.0418

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：

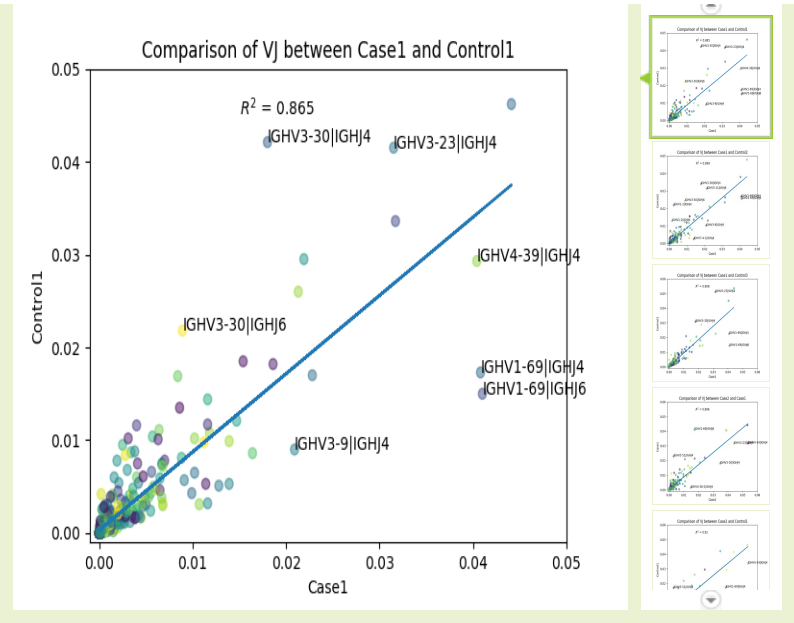

图 2-11 样本间VJ基因共线性

表 2-5 样本间VJ基因组合的频率差异

VJGene	Case3	Case2	Case1	Control3	Control1	Control2
IGHV3-11|IGHJ4	0.0441	0.0533	0.0493	0.0462	0.0479	0.0536
IGHV1-69|IGHJ6	0.041	0.0173	0.0203	0.015	0.0254	0.0143
IGHV1-69|IGHJ4	0.0408	0.0385	0.027	0.0173	0.0267	0.0225
IGHV4-39|IGHJ4	0.0404	0.0391	0.0327	0.0293	0.038	0.0451
IGHV4-34|IGHJ4	0.0317	0.0536	0.0326	0.0336	0.0263	0.0226
IGHV3-23|IGHJ4	0.0315	0.044	0.0495	0.0415	0.0235	0.0511
IGHV5-51|IGHJ4	0.0228	0.0031	0.0087	0.017	0.021	0.0146
IGHV3-48|IGHJ4	0.0219	0.0243	0.0267	0.0295	0.0131	0.0288
IGHV3-11|IGHJ6	0.0213	0.0174	0.0184	0.026	0.0313	0.018
IGHV3-9|IGHJ4	0.0209	0.0108	0.0182	0.009	0.0102	0.0143
IGHV3-15|IGHJ4	0.0186	0.021	0.0172	0.0182	0.0143	0.0205
IGHV3-30|IGHJ4	0.018	0.0348	0.0222	0.0421	0.034	0.0308
IGHV2-5|IGHJ4	0.0164	0.013	0.0088	0.0086	0.0157	0.0149
IGHV4-34|IGHJ6	0.0154	0.0155	0.0196	0.0185	0.0152	0.0101

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：

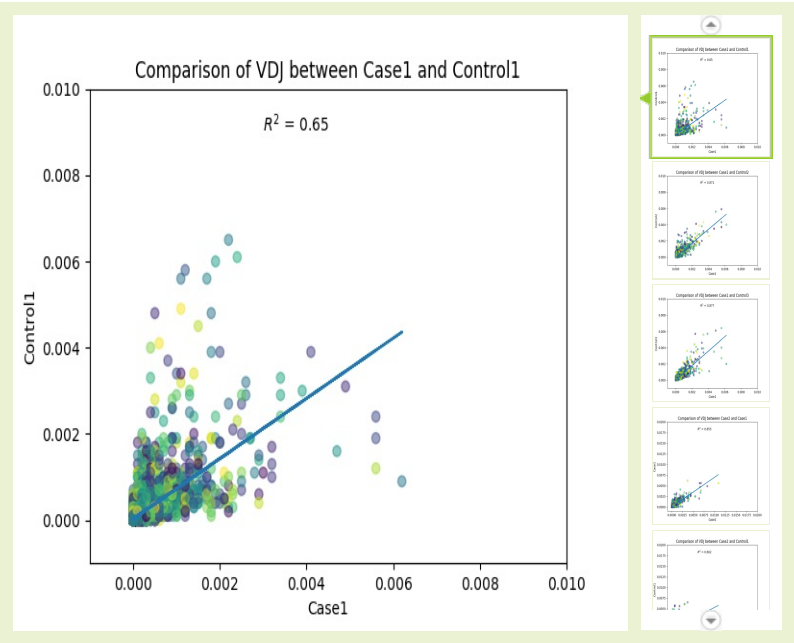

图 2-12 样本间VDJ基因共线性

表 2-6 样本间VDJ基因组合的频率差异

VDJGene	Case3	Case2	Case1	Control3	Control1	Control2
IGHV1-69|IGHD3-10|IGHJ6	0.0062	0.0037	0.0017	0.0009	0.004	0.002
IGHV1-69|IGHD3-10|IGHJ4	0.0056	0.0065	0.0021	0.0012	0.0037	0.0027
IGHV4-34|IGHD3-10|IGHJ4	0.0056	0.0109	0.0029	0.0024	0.0043	0.004
IGHV4-39|IGHD3-10|IGHJ4	0.0056	0.0068	0.0027	0.0019	0.0059	0.0064
IGHV3-11|IGHD3-10|IGHJ4	0.0049	0.0083	0.0037	0.0031	0.0056	0.0061
IGHV1-69|IGHD3-22|IGHJ4	0.0047	0.0066	0.0037	0.0016	0.0035	0.0035
IGHV1-69|IGHD2-2|IGHJ6	0.0041	0.002	0.0034	0.0039	0.0026	0.0011
IGHV3-23|IGHD3-10|IGHJ4	0.0039	0.0068	0.0034	0.003	0.0031	0.0057
IGHV3-11|IGHD6-13|IGHJ4	0.0034	0.004	0.0032	0.0033	0.0032	0.0039
IGHV4-39|IGHD6-13|IGHJ4	0.0034	0.0029	0.0023	0.0024	0.003	0.0034
IGHV3-11|IGHD3-22|IGHJ4	0.0034	0.0067	0.0052	0.0029	0.0043	0.0056
IGHV4-39|IGHD3-22|IGHJ4	0.0032	0.0045	0.0035	0.0017	0.0038	0.0046
IGHV4-39|IGHD6-19|IGHJ4	0.0032	0.0034	0.0028	0.0013	0.0029	0.0039
IGHV1-69|IGHD6-13|IGHJ6	0.0032	0.0011	0.0014	0.001	0.002	0.0012

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：

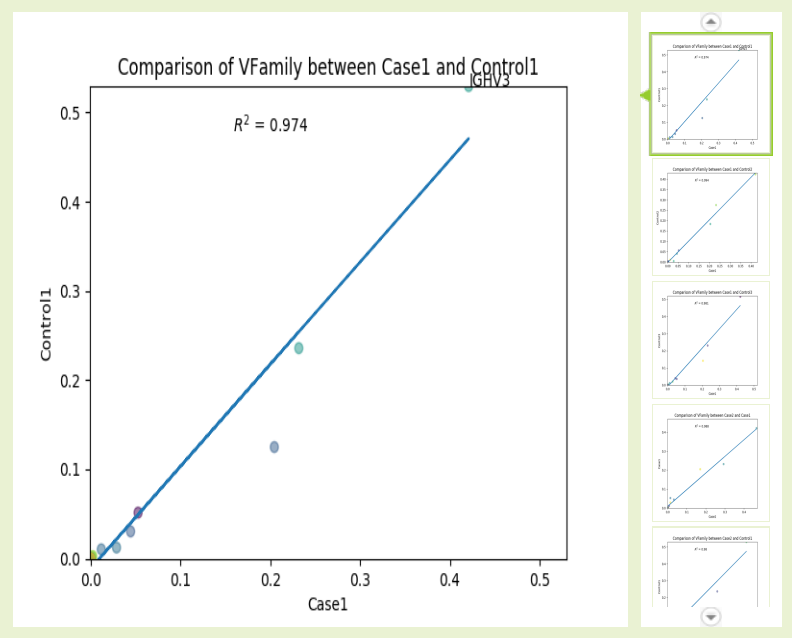

图 2-13 样本间VFamily基因共线性

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Clonetype频率进行假设检验分析，部分结果如下表所示：

表 2-6 样本间假设检验分析

Clonetype	Case3	Case2	Case1	Control3	Control1	Control2	Pvalue	Qvalue
CARGAYYFDYW	8.58798585763e-06	9.25613769348e-06	1.47415514125e-05	0	2.41523783209e-07	1.64994910732e-06	0.00710679732272	0.00452209294837
CARGEGWFDPW	8.17903415013e-06	9.46182964223e-06	1.34279478213e-05	0	0	6.76479134001e-06	0.0422873263301	0.00452209294837
CARDPPDYW	7.15665488136e-06	9.05044574474e-06	5.54632627403e-06	0	0	0	0.00201331705336	0.00452209294837
CARIGYSSSCFDYW	4.90742049008e-06	4.93660676986e-06	7.29779772898e-06	1.65582816743e-07	0	1.64994910732e-07	0.00212464157571	0.00452209294837
CAREYSSSSGRAFDIW	4.70294463632e-06	9.87321353971e-06	8.17353345646e-06	1.65582816743e-07	0	1.64994910732e-07	0.00799430825585	0.00452209294837
CATDLLDYW	3.68056536756e-06	3.2910711799e-06	6.42206200151e-06	0	2.41523783209e-06	0	0.0451818764885	0.00452209294837
CVKGGWLDDW	2.04475853753e-06	3.90814702614e-06	1.60551550038e-06	1.65582816743e-07	0	3.29989821464e-07	0.0297779066573	0.00452209294837
CASGAYW	2.04475853753e-	2.05691948744e-	1.4595595458e-	0	0	0	0.000712626264066	0.00452209294837

	07	07	07					
CARGGGNFDYW	2.04475853753e-07	4.31953092362e-06	1.4595595458e-07	1.90420239254e-05	1.18346653773e-05	8.41474044733e-06	0.0280126480716	0.00452209294837
CARGDYW	0	0	0	6.12656421947e-06	7.00418971307e-06	1.22096233942e-05	0.0112516272524	0.00452209294837

2.7.3 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Clonetype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Clonetype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：

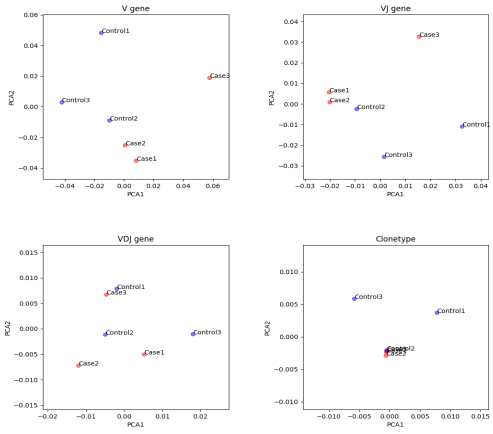

图 2-14 样本间PCA分析(点击大图)

三、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] 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.
[6] Jennifer Benichou, Rotem Ben-Hamo, Yoram Louzoun et al. 2Rep-Seq uncovering the immunological repertoire through next-generation sequencing. Immunology, 2011, 135: 183–191.
[7] 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.
[8] 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. OncoImmunology, 2015, Vol.4:e1030561-1-10.
[9] 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.
[10] X-L Hou, LWang, Y-L Ding, et al. Current status and recent advances of next generation sequencing techniques in immunological repertoire. Genes and Immunity, 2016, 1–12. [11] 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.

四、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、ClonetypeFraction.txt：样本间clonetype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
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/