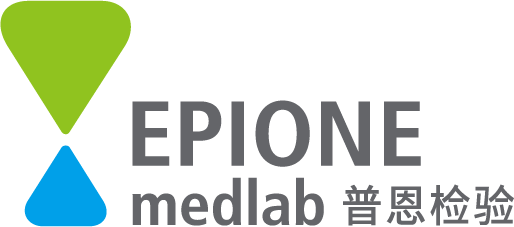
TCR测序结题报告



课题负责人：刘主任

客户单位：中山医院

# 简介

T细胞受体(T cell receptor，TCR)是T细胞表面特异性识别抗原和介导免疫应答的分子，是人类基因组中多态性最高的区域之一，决定着人的免疫系统如何适应环境变化。

TCR是高度多样化的异源二聚体, 可分为TCRα/β和TCRγ/δ两种类型，外周血的TCR以αβ组合为主。TCRβ基因由可变区(V)、多变区(D)、 结合区(J)和恒定区(C)四种基因片段组成，四种基因片段组合形成互补决定区 (complementarities determining region，CDR：CDR1、CDR2、CDR3）和间隔的4 个骨架区(framework region，FR)。其中，CDR3区由V区、D区和J区重排形成具有功能的TCR编码基因。

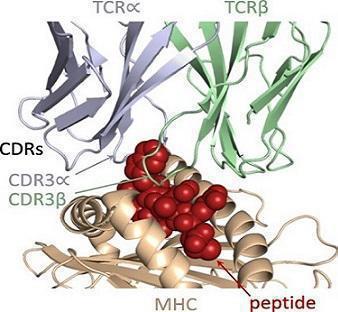


图1 TCRα/β与MHC及抗原肽的互作

在T细胞发育过程中，CDR1和CDR2由V基因编码，对TCR与MHC复合物互作至关重要。CDR1，CDR2和FR区域相对保守；而CDR3由于V (65-100种)、D (2种)、J (13种)基因片段本身具有的多样性，以及基因重排过程中 V-D及D-J连接区存在的非模板核苷酸随机插入或删除，使得CDR3区多样性非常丰富。除非T细胞是来源于相同的克隆，一般来说T细胞几乎不太可能表达相同的CDR3序列。正是因为这种丰富的多样性，机体可以识别各种不同的抗原。

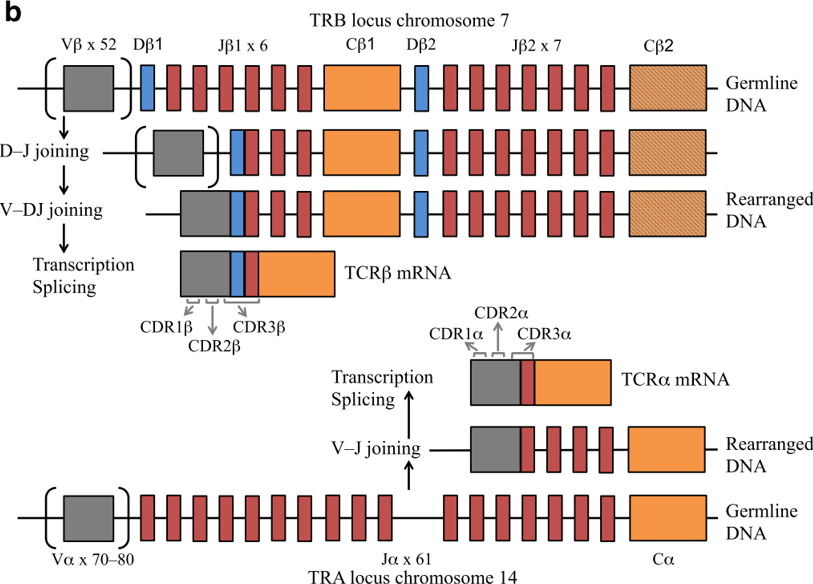


图2 TCRα/β的V(D)J 基因重组示意图

以T淋巴细胞为研究目标，通过多重PCR技术扩增T细胞受体(TCRβ链)的CDR3区，结合深度测序技术，可以对机体TCR的多样性进行全面评估，使得深入挖掘免疫系统与疾病的关系成为可能。

# 实验流程

本项目采用赛默飞TCR免疫组库检测NGS试剂盒（Oncomine TCR Beta-LR Assay），实现全面的TCR beta测序，获得包括TRBV等位基因在内的分型结果，用于免疫应答相关的克隆型分析以及工程化T细胞监测。该试剂盒兼容血液、新鲜冻存组织、分选细胞， 并以双条码技术进一步降低引物干扰和误差，提升检测灵敏度。实验流程概括如下：

（1）提取RNA 使用特定引物反转成 cDNA。

（2）将样本中加入配对引物 ，进行靶标扩增PCR 反应。如下图：

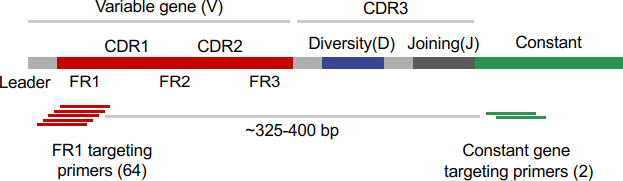


图3 Oncomine™ TCR Beta‑LR Assay 多重PCR设计示意图.

（3）将扩增产物酶切消化后进行连接反应，引入接头序列和测序引物。

（4）采用Agencourt AMPure XP 磁珠纯化文库。

（5）文库质检：检测文库片段范围和浓度定量。

（6）上机测序

实验流程图如下：

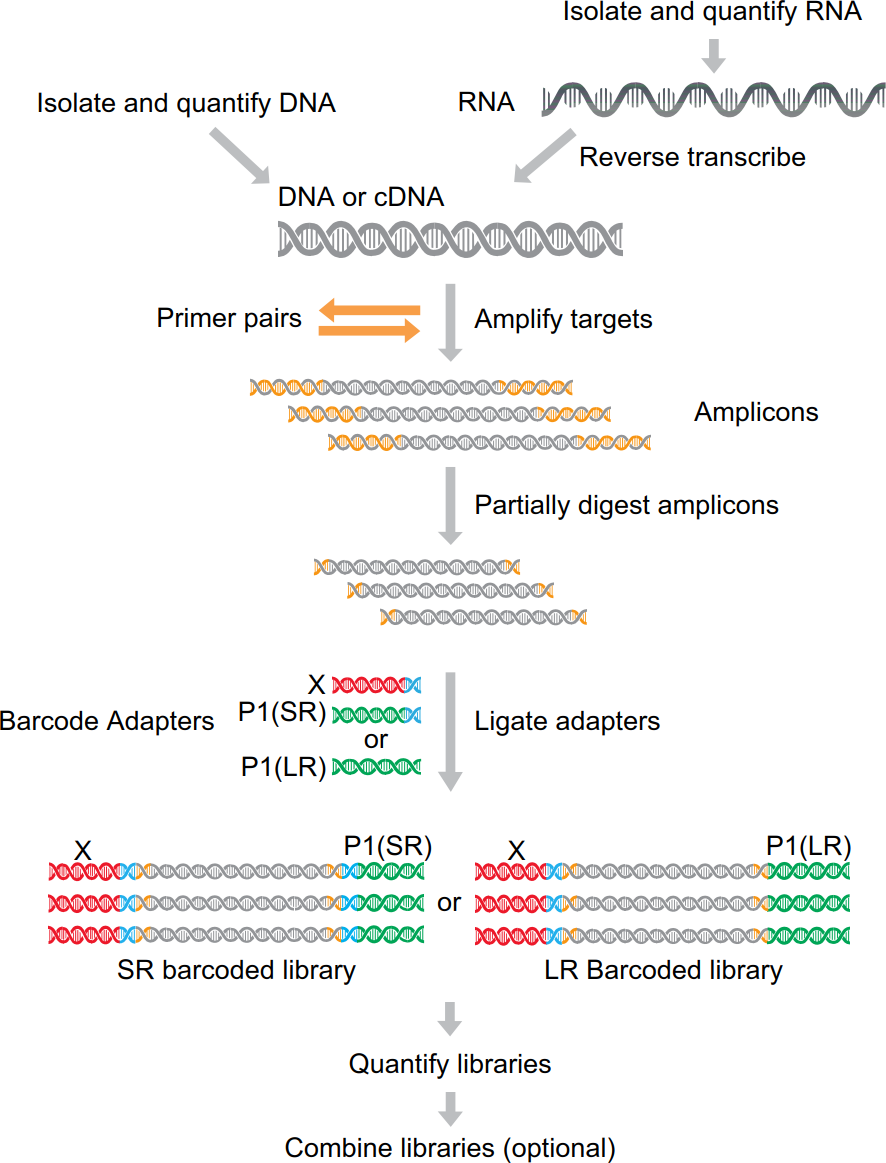


图4 实验流程图

# 分析流程

（1）获得下机测序原始数据；（2）经过滤得到clean reads并比对到参考序列； （3）根据比对结果得到具体的功能区域，例如CDR3区（clones）；（4）碱基质量符合要求的克隆序列会作为核心克隆，存在⼀个以上质量值较差碱基的克隆会以核心克隆作参考⼆次比对和校正；（5）对相差⼀个碱基的克隆，进行层次聚类，每个分支间仅有⼀个碱基差别（mismatch），依次聚类下去，克隆频率低的克隆会合并到上⼀分支，最终保留最顶端序列；（6）将上述得到的克隆序列再次比对到V，D，J和C参考序列， 最终统计克隆序列、氨基酸残基序列、克隆数量、克隆频率、V/J基因组合等信息，后续可以根据这些信息做克隆分布、多样性等深⼊分析。参考序列来自国际公共数据库 IMGT (http://www.imgt.org/)。

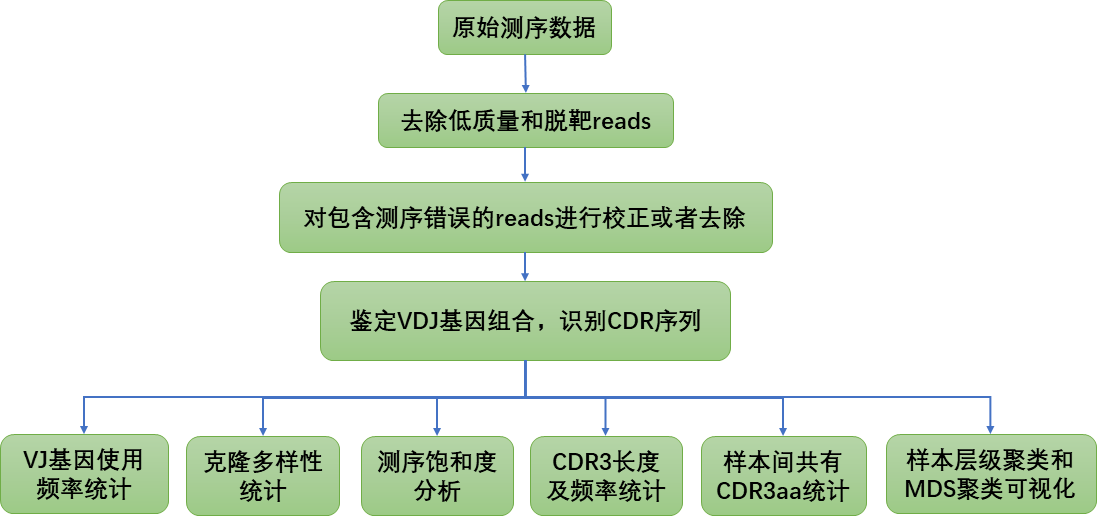


图5 生信分析流程图

# 样本信息

样本信息归纳如下表，包含了样本名称，样本分组，样本描述等信息

表1 样本信息表

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 患者编号 | 病人代号 | 样本编号 | 分析编号 | Organ | tumour | 临床诊断 |
| EP19A0195 | 845661848 | ZD904232007 | nasopharynx\_01 | nasopharynx | 鼻咽恶性肿瘤 | 鼻咽恶性肿瘤 |
| EP19A0198 | 846121043 | ZD904232010 | nasopharynx\_02 | nasopharynx | 鼻咽恶性肿瘤 | 鼻咽恶性肿瘤个人史 |
| EP19A0380 | 1060961 | ZD906232001 | nasopharynx\_03 | nasopharynx | 鼻咽恶性肿瘤 | 鼻咽恶性肿瘤 |
| EP19A0131 | 1077720 | ZD903232003 | lung\_01 | lungs | 肺恶性肿瘤 | 支气管或肺恶性肿瘤(腺癌 cT2aN0M1a IV期 肺转移 T3胸椎转移) |
| EP19A0171 | 1136132 | ZD904232002 | lung\_02 | lungs | 肺恶性肿瘤 | 支气管或肺恶性肿瘤 |
| EP19A0196 | 1195628 | ZD904232008 | lung\_03 | lungs | 肺恶性肿瘤 | 支气管或肺恶性肿瘤 |
| EP19A0170 | 1118523 | ZD904232001 | billiary\_tract\_01 | billiary\_tract | 肝内胆管癌 | 肝内胆管癌(术后 IV期) |
| EP19A0337 | 1203364 | ZD905232001 | cervix\_01 | cervix | 宫颈恶性肿瘤 | 宫颈恶性肿瘤 |
| EP19A0381 | 1210549 | ZD906232002 | maxilla\_01 | maxilla | 上颌骨恶性肿瘤 | 上颌骨恶性肿瘤 |
| EP19A0174 | 890204 | ZD904232005 | oesophagus\_01 | oesophagus | 食管恶性肿瘤 | 食管恶性肿瘤 |
| EP19A0169 | 845838048 | ZD903232004 | stomach\_01 | stomach | 胃恶性肿瘤 | 胃恶性肿瘤伴肝转移 |
| EP19A0173 | 845958788 | ZD904232004 | stomach\_02 | stomach | 胃恶性肿瘤 | 胃恶性肿瘤 |
| EP19A0175 | 845960906 | ZD904232006 | stomach\_03 | stomach | 胃恶性肿瘤 | 胃恶性肿瘤 |
| EP19A0338 | 1167192 | ZD905232002 | stomach\_04 | stomach | 胃恶性肿瘤 | 胃恶性肿瘤（HER2阴性型）(ypT4N1M0 IIIB期) |
| EP19A0199 | 846132081 | ZD904232011 | hypopharynx\_01 | hypopharynx | 下咽恶性肿瘤 | 下咽恶性肿瘤 |
| EP19A0172 | 845935005 | ZD904232003 | melanoma\_01 | melanoma | 眼睑恶性黑色素瘤 | 眼睑恶性黑色素瘤 |

# 质控结果

TCR变化多端，要尽可能多地获得不同克隆T细胞的CDR3的序列，需要有足够的测序深度；CDR序列之间有时仅相差一个碱基，因此要对原始测序reads进行严格质控，对于PCR错误还要尽可能进行校正。如下表，我们统计了测序reads的数量和测序捕获信息等。

表2 质控信息表

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| SampleID | proportion\_germline\_reads | proportion\_perfect\_reads | Productive\_reads | Rescued\_productive\_reads | Unproductive\_reads | Off\_target\_reads |
| ZD904232007 | 0.983 | 0.838 | 255489 | 339272 | 167587 | 223250 |
| ZD904232010 | 0.992 | 0.892 | 250393 | 356497 | 176722 | 228579 |
| ZD906232001 | 0.992 | 0.869 | 257342 | 339415 | 151351 | 205428 |
| ZD903232003 | 0.984 | 0.82 | 274090 | 349646 | 178438 | 175423 |
| ZD904232002 | 0.988 | 0.819 | 282258 | 363981 | 166691 | 206972 |
| ZD904232008 | 0.99 | 0.872 | 237304 | 311408 | 146745 | 219756 |
| ZD904232001 | 0.989 | 0.837 | 205971 | 300605 | 146436 | 221067 |
| ZD905232001 | 0.991 | 0.886 | 282162 | 360645 | 164121 | 205784 |
| ZD906232002 | 0.991 | 0.869 | 290048 | 356481 | 164795 | 219944 |
| ZD904232005 | 0.989 | 0.859 | 281800 | 367732 | 193427 | 244895 |
| ZD903232004 | 0.981 | 0.83 | 258635 | 338823 | 149957 | 212718 |
| ZD904232004 | 0.99 | 0.889 | 272884 | 386442 | 171533 | 230113 |
| ZD904232006 | 0.992 | 0.867 | 308339 | 392257 | 163119 | 180376 |
| ZD905232002 | 0.99 | 0.865 | 234585 | 291402 | 134414 | 307951 |
| ZD904232011 | 0.992 | 0.927 | 280314 | 361141 | 166242 | 230378 |
| ZD904232003 | 0.99 | 0.877 | 312495 | 406141 | 258065 | 252561 |

[Note] Productive reads：有效reads，包含V和J基因序列但不包含终止子的理论可编码reads； Rescued productive reads：经过错误校正后得到的有效reads； Proportion germline reads：种系reads比例，即和参考V基因的碱基错配率小于1%的有效reads，包括经过校正得到的有效reads； Proportion perfect reads：完美reads比例，和参考V基因的碱基完全匹配的有效reads，包括经过校正得到的有效reads； Unproductive reads：无效reads，和有效reads相对，包含不可校正的测序错误或pcr错误，导致无法编码或提前翻译终止。

# 分析结果

## 克隆信息总表

通过CDR序列分析，把CDR3aa序列作为克隆的唯一标识，对每个样本的T细胞克隆多样性进行统计，汇总信息如下表，具体表格见{{分析结果文件名}}。

表3 克隆多样性统计表

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| SampleID | NewName | Clones | Evenness | Reads | Shannon\_diversity | convergent\_TCR\_frequency | clone\_gini\_index |
| ZD904232007 | nasopharynx\_01 | 8726 | 0.799 | 227398 | 10.462 | 0.019 | 0.707 |
| ZD904232010 | nasopharynx\_02 | 6543 | 0.745 | 308083 | 9.442 | 0.008 | 0.751 |
| ZD906232001 | nasopharynx\_03 | 6088 | 0.906 | 301843 | 11.393 | 0.014 | 0.588 |
| ZD903232003 | lung\_01 | 8141 | 0.582 | 197414 | 7.558 | 0.007 | 0.834 |
| ZD904232002 | lung\_02 | 9876 | 0.809 | 264145 | 10.735 | 0.013 | 0.707 |
| ZD904232008 | lung\_03 | 7208 | 0.746 | 206924 | 9.557 | 0.019 | 0.791 |
| ZD904232001 | billiary\_tract\_01 | 4697 | 0.62 | 240795 | 7.566 | 0.015 | 0.843 |
| ZD905232001 | cervix\_01 | 6639 | 0.768 | 323686 | 9.754 | 0.005 | 0.725 |
| ZD906232002 | maxilla\_01 | 9626 | 0.819 | 289038 | 10.842 | 0.011 | 0.689 |
| ZD904232005 | oesophagus\_01 | 11744 | 0.856 | 212024 | 11.577 | 0.023 | 0.644 |
| ZD903232004 | stomach\_01 | 7071 | 0.781 | 256199 | 9.981 | 0.083 | 0.7 |
| ZD904232004 | stomach\_02 | 10252 | 0.915 | 275991 | 12.188 | 0.011 | 0.578 |
| ZD904232006 | stomach\_03 | 5949 | 0.552 | 394910 | 6.925 | 0.112 | 0.885 |
| ZD905232002 | stomach\_04 | 9966 | 0.857 | 192309 | 11.387 | 0.009 | 0.64 |
| ZD904232011 | hypopharynx\_01 | 7717 | 0.849 | 323179 | 10.962 | 0.048 | 0.678 |
| ZD904232003 | melanoma\_01 | 12098 | 0.899 | 292949 | 12.191 | 0.021 | 0.588 |

[Note] Clones：克隆类型数目； Shannon\_diversity：香浓多样性指数，根据克隆频率计算，该值越高，表示克隆多样性越高； Evenness：标准化后的香浓多样性指数=Shannon\_diversity/log2(Clones)；Reads：最后用于鉴定出CDR的有效reads；convergent\_TCR\_frequency：CDR3核酸序列不一致但CDR3的氨基酸序列一致，这种情况称之为TCR convergence, 把这种克隆的占比称为convergent TCR frequency，该指标可以一定程度反映免疫响应效果； clone\_gini\_index：基尼指数，类似香浓多样性，该值越大，表示克隆多样性越高。

除了克隆多样性汇总表，针对每个样本，我们提供了每个克隆的详细信息，包括VDJ组合，CDR3序列信息等，后续分析均基于这些信息。单个样本克隆信息汇总示例如下表，具体表格见{{分析结果文件名}}。

表4 单个样本克隆信息总表

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Variable | Joining | CDR3 AA | CDR3 NT | Diversity | ... | Diversity Gene Allele | Joining Gene Allele | Constant Gene | Sequence | Reference Sequence |
| TRBV15 | TRBJ2-3 | ATSRDTLA... | GCCACCAG... | TRBD2 | ... | 2.0 | 1 | TRBC2 | GTTCTCAG... | GTTCTCAG... |
| TRBV6-5 | TRBJ1-1 | ASSYKIGWNTEAF | GCCAGCAG... | nan | ... | nan | 1 | TRBC1 | GAAGACAG... | GAAGACAG... |
| TRBV29-1 | TRBJ2-7 | SVLTGVYEQY | AGCGTCCT... | TRBD1 | ... | 1.0 | 1 | TRBC2 | CAACGTGG... | CAACGTGG... |
| TRBV6-2 | TRBJ2-2 | ASSSLDTGELF | GCCAGCAG... | nan | ... | nan | 1 | TRBC2 | GAAGACAG... | GAAGACAG... |
| TRBV9 | TRBJ1-6 | ASSPSNSPLH | GCCAGCAG... | nan | ... | nan | 1 | TRBC1 | GGACAGCG... | GGACAGCG... |
| TRBV15 | TRBJ2-3 | ATSRDTLA... | GCCACCAG... | TRBD2 | ... | 1.0 | 1 | TRBC2 | GTTCTCAG... | GTTCTCAG... |
| TRBV29-1 | TRBJ1-1 | SVDRTEAF | AGCGTTGA... | nan | ... | nan | 1 | TRBC1 | CAACGTGG... | CAACGTGG... |
| TRBV6-2 | TRBJ2-4 | ASATSSQGIQY | GCCAGCGC... | TRBD1 | ... | 1.0 | 1 | TRBC2 | GAAGACAG... | GAAGACAG... |
| TRBV28 | TRBJ2-1 | ASSLRDLAGYNEQF | GCCAGCAG... | TRBD2 | ... | 1.0 | 1 | TRBC2 | CAAAAGGA... | CAAAAGGA... |
| TRBV2 | TRBJ2-1 | ASRLRDESYNEQF | GCCAGCAG... | TRBD2 | ... | 1.0 | 1 | TRBC2 | CACACAGA... | CACACAGA... |
| TRBV5-1 | TRBJ1-1 | ASKGVNTEAF | GCCAGTAA... | TRBD1 | ... | 1.0 | 1 | TRBC1 | CTGATCAA... | CTGATCAA... |
| TRBV20-1 | TRBJ2-3 | SATFGTDTQY | AGTGCCAC... | nan | ... | nan | 1 | TRBC2 | GGGTTATC... | GGGTTATC... |

[Note] Variable：V基因id； Joining：J基因id； CDR3 AA：CDR3氨基酸序列； CDR3 NT：CDR3核酸序列； Diversity：D基因id； Plus Counts：比对到正义链的read数量； Minus Counts：比对到负链的read数量； Variable Mutation：V基因碱基突变率； Total Counts：属于当前clone的总read数量； Frequency：当前克隆的出现频率； Rank：根据Frequency的排序得到的排名； Variable Gene Allele：V基因等位基因数目； Diversity Gene Allele：D等位基因数目； Joining Gene Allele：J等位基因数目； Constant Gene：C基因id； Sequence：本次测序鉴定出的VDJ组合序列 Reference Sequence：参考基因序列

## 测序饱和度分析

测序饱和度分析可以在⼀定程度上判断测序数据量是否满足要求。这里用到的方法是逐步扩大随机抽样的容量，随着容量即测序量的增加，检测到的克隆数也随之上升，当测序量达到⼀定区间后其检测到的克隆数量增加变缓，说明此时测序量达到要求。注意，实际情况往往是随着测序量增加，克隆数也一直增加，但是这并非意味着测序量越高越好，因为通过高深度测序得到的克隆往往是稀有clone，对后续分析可能没有任何帮助，甚至增加测序错误或偏差。

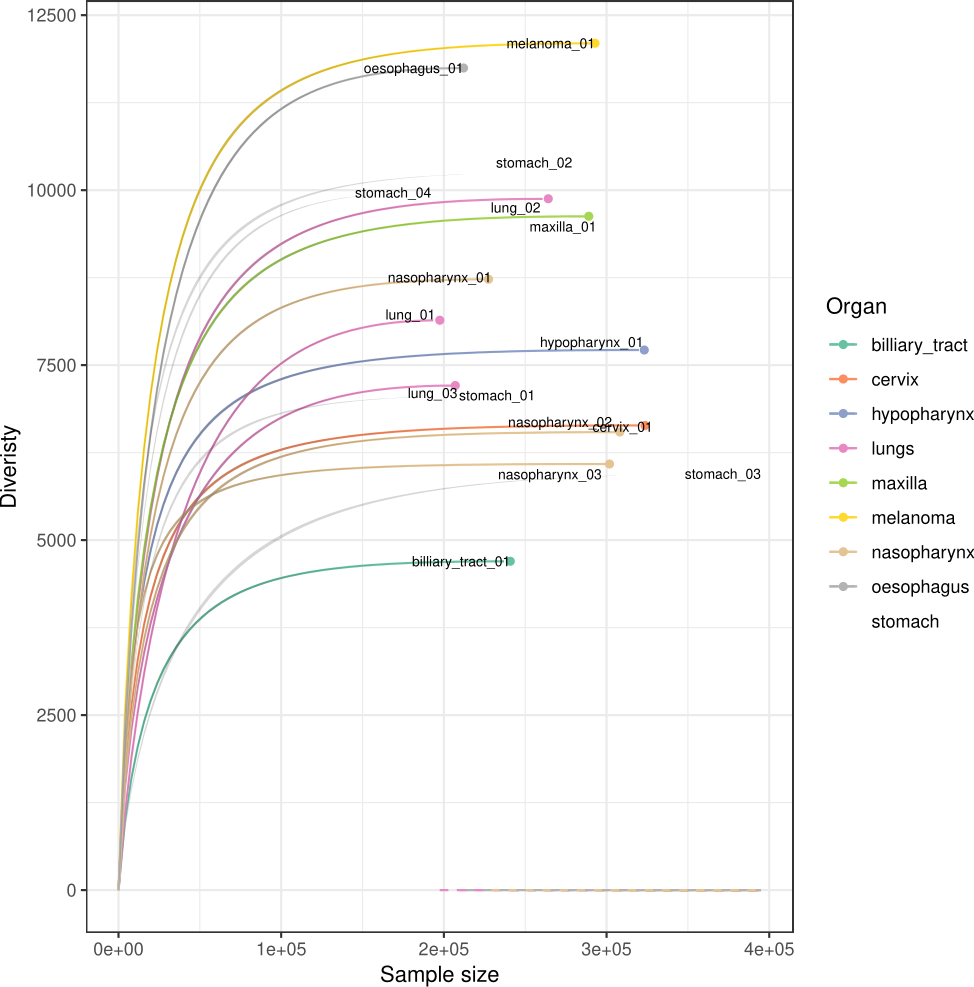


图6 测序饱和度分析 横轴表示样本测序量，纵轴表示对应的克隆数大小，实线表示抽样量低于实际测序量，虚线表示抽样量高于实际测序量，圆点表示此时的抽样量等于实际测序量，阴影区域表示95%的置信区间。

## V基因使用频率统计

V基因是TCR中可变性最多的基因，许多研究表明V基因的使用情况可以反映机体对病菌、病毒、肿瘤抗原的免疫响应变化，如有些病人在治疗前看不到特定的V基因使用偏好，但是治疗起效后，有些V基因的使用出现了偏好。我们基于V基因的使用频率矩阵绘制了如下聚类热图。

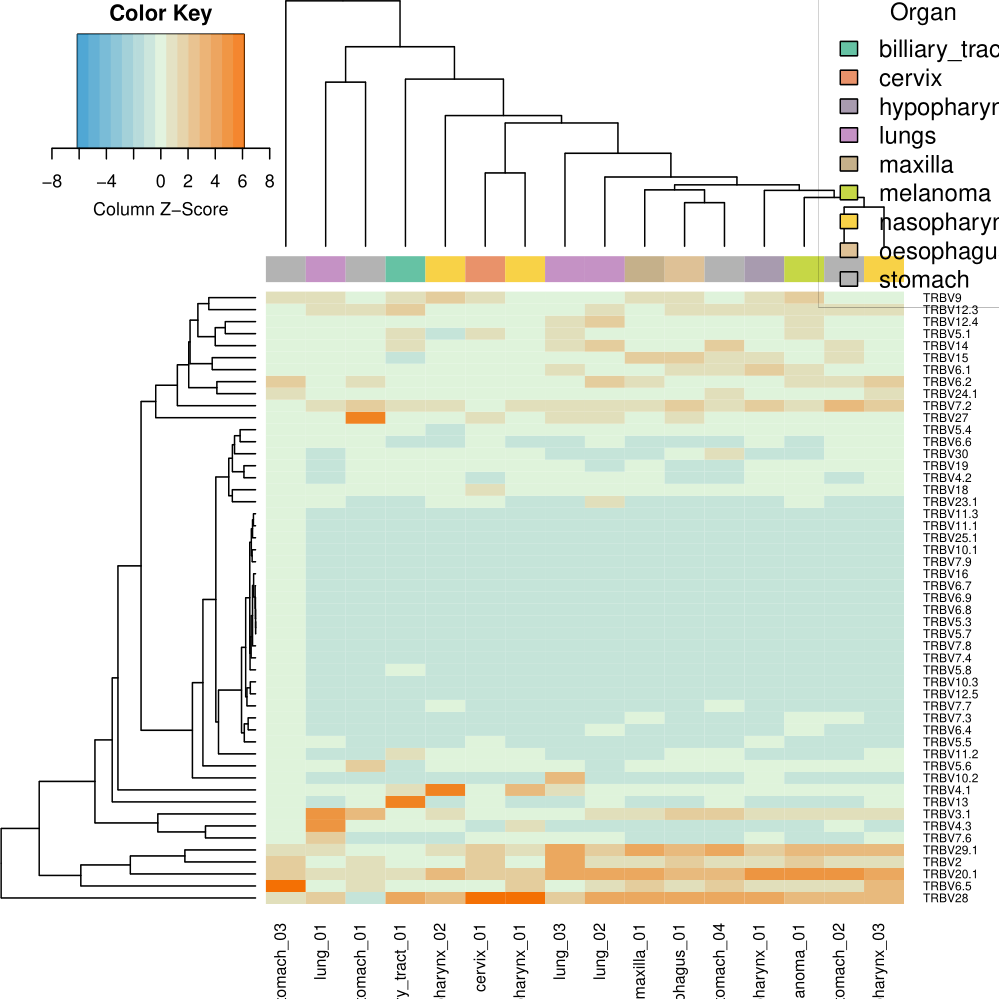


图7 V基因使用情况分析 热图颜色由冷色过渡到暖色，表示使用频率从低到高。图上方为样本聚类树，左侧为基因聚类树，在所有样本中，基因使用率相似的基因聚类到一起。

## J基因使用频率统计

J基因 也是TCR中可变性较多的基因，其和V基因的重组可以构成CDR3区域。我们基于J基因的使用频率矩阵绘制了如下聚类热图。

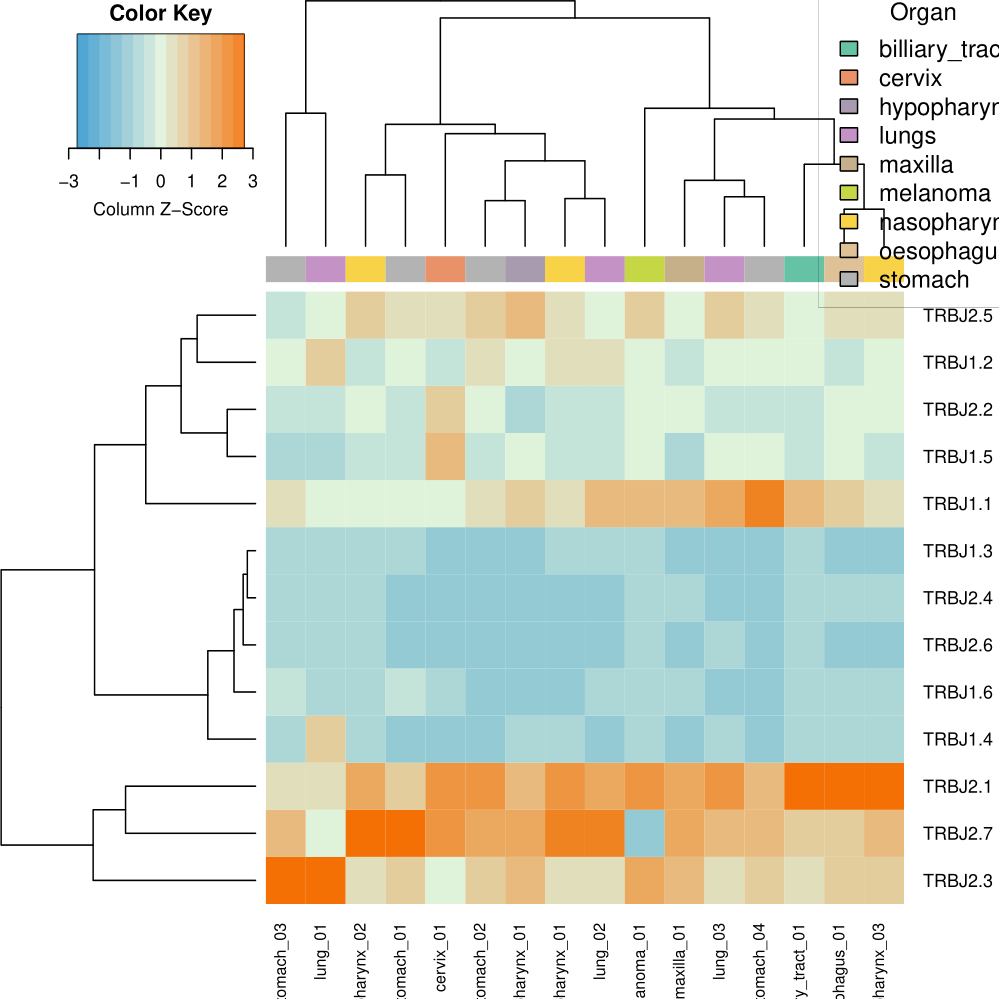


图8 J Usage. 热图颜色由冷色过渡到暖色，表示使用频率从低到高。图上方为样本聚类树，左侧为基因聚类树，在所有样本中，基因使用率相似的基因聚类到一起。

## V-J配对使用频率统计

V和J基因也是CDR3区域不可或缺的部分，而CDR3几乎决定了其T细胞克隆来源的唯一性。针对每个样本，我们基于V、J使用频数绘制了如下弦图：

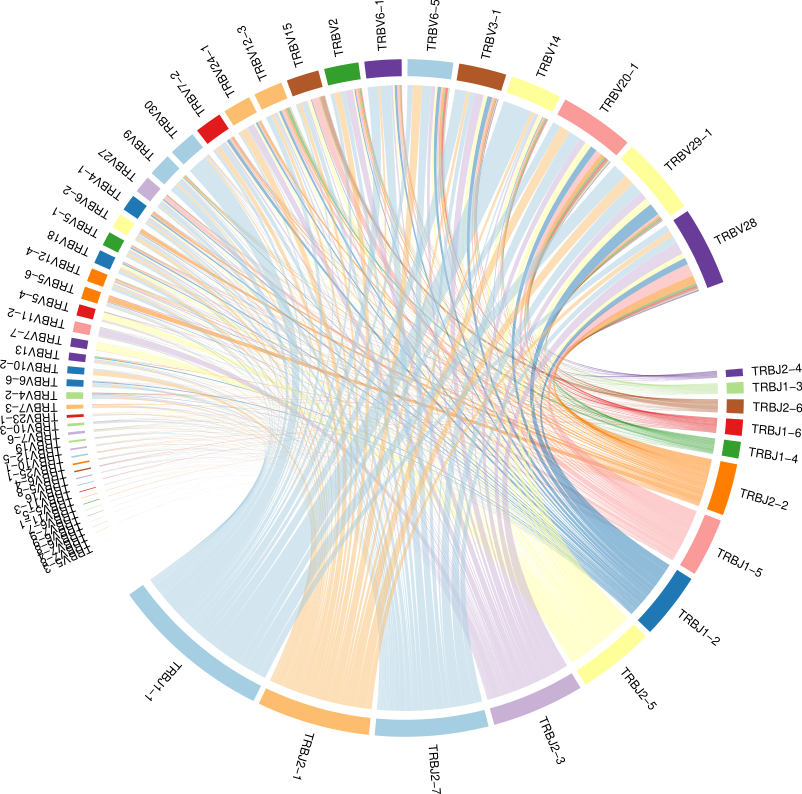


图9 V-J基因配对统计弦图 样本：stomach\_04。上弦弧包含的是V基因，下弦弧包含的是J基因，不同基因用不同的颜色区分，弧线越长则表示使用频数越多，每根弦连接一对VJ基因。

除了弦图，我们还可以使用立体的柱状图来直观地展示VJ基因的配对使用频率高低，如下图：

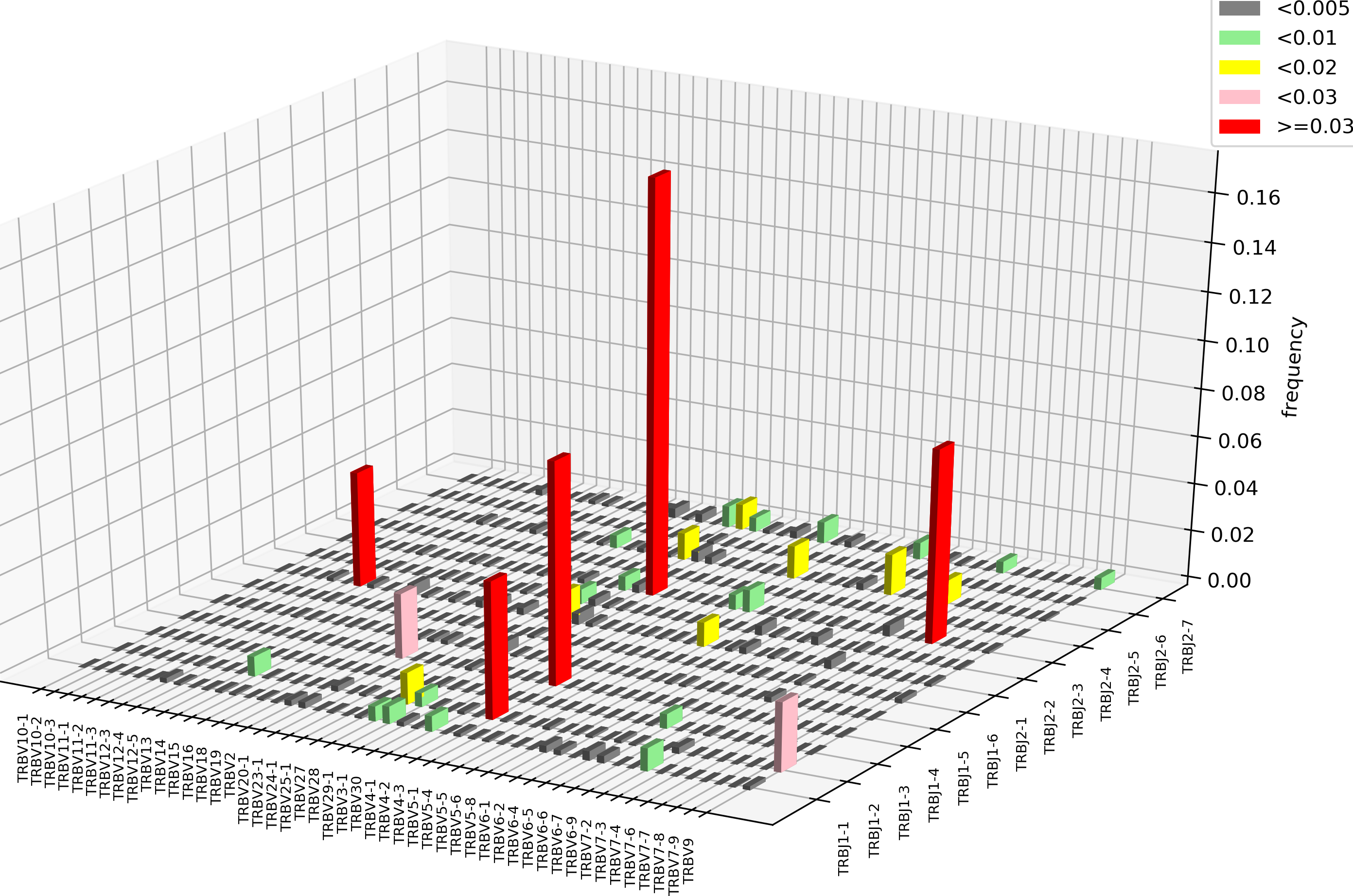


图10 V-J基因配对统计柱图 样本：lung\_01。

## CRD3长度分布统计

CDR3是和抗原肽直接互作的区域，对于抗原的识别起着至关重要的作用。由于CDR3是V(D)J组合的结果，其多样性以数百亿计，这几乎导致一个CDR3只可能来源于相同的T细胞克隆。针对每个样本，我们对CDR3的核酸序列长度进行分布统计，并标注了部分克隆数排名靠前的克隆类型，用CDR3的氨基酸序列表示。

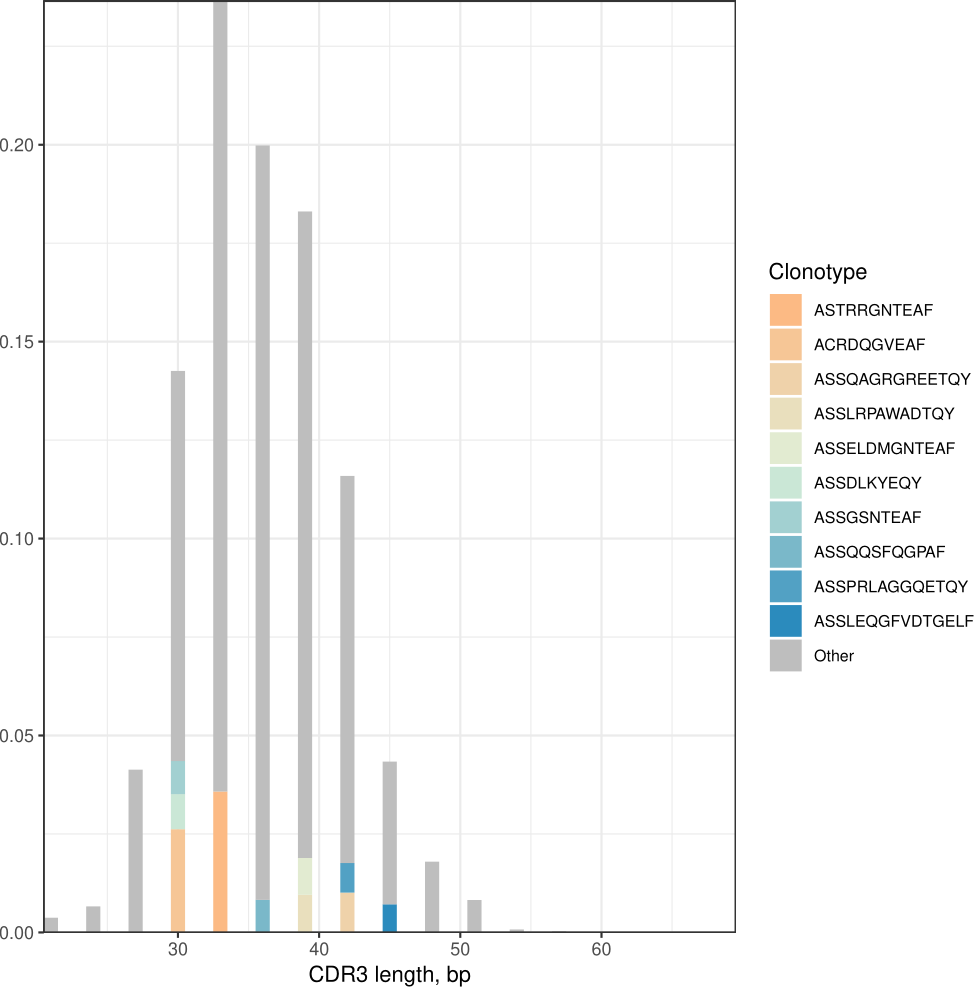


图11 CRD3核酸长度分布统计。样本：stomach\_04。横轴表示核酸序列长度，纵轴表示某个特定核酸序列长度的CDR3的频率。彩色柱块标注频率排名靠前的克隆类型，以氨基酸序列长度表示。

## 共有克隆统计

用CDR3的氨基酸序列作为克隆的唯一标识，我们可以对多个样本间的共有克隆或CDR3aa进行统计，如下图，以Venn图的形式呈现。共有克隆越多，一定程度反映样本越相似。

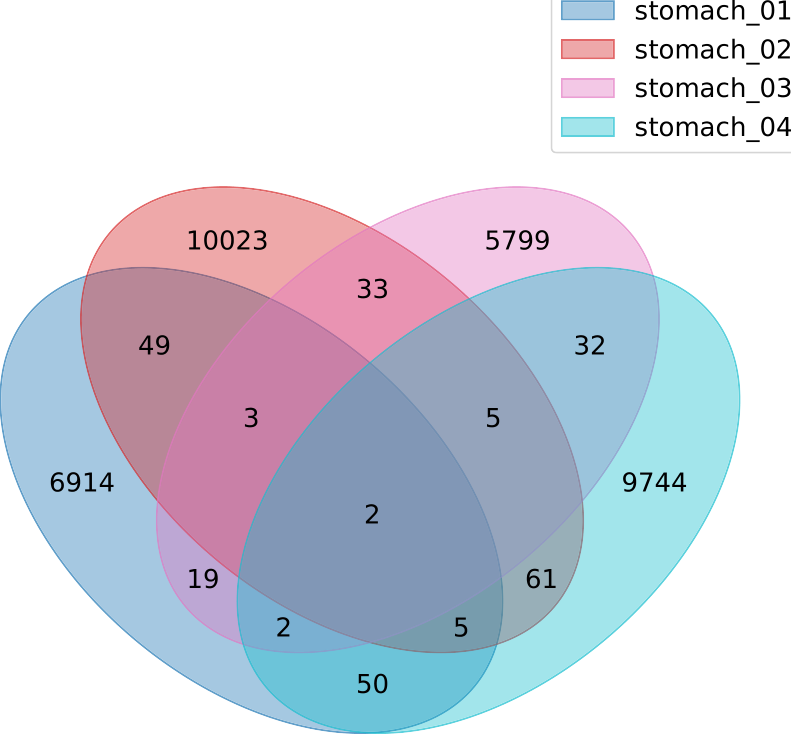


图12 共有克隆分析 组别：stomach。不同颜色的圆表示不同的样本。

## 样本聚类

衡量样品间相似性的指标有多种，没有黄金标准。主要包括克隆频率Pearson相关系数（P），相对克隆重叠 （D），克隆累积频率⼏何平均数（F），Сlonotype-wise克隆重叠⼏何距离（F2），Morisita-Horm index等，详见VDJtools文档(https://vdjtools-doc.readthedocs.io/en/master/overlap.html#calcpairwisedistances)。我们采用vdjtools默认的“克隆累积频率⼏何平均数”作为样本间距离度量进行层级聚类，如下图。克隆累积频率⼏何平均数 = 共有克隆分别在两个样本中的频率总和的乘积的开方。

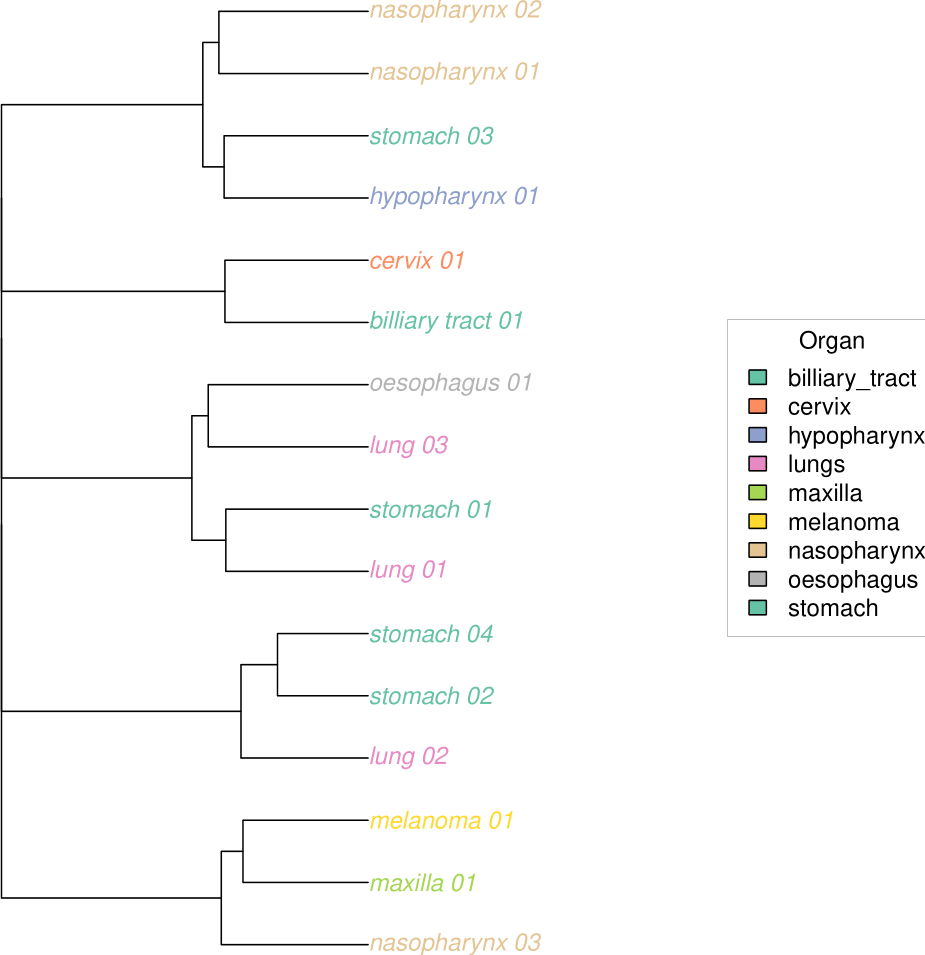


图13 样本层级聚类树。分支干线长度表示了两个样本间的距离，不同字体颜色区分不同的样本分组。

基于样本间距离矩阵，不仅可以进行层级聚类，还可以进行MDS(multidimensional scaling)多维尺度分析。MDS是一种降维方法，方便从较低的维度如二维去看样本间的距离，如下图。

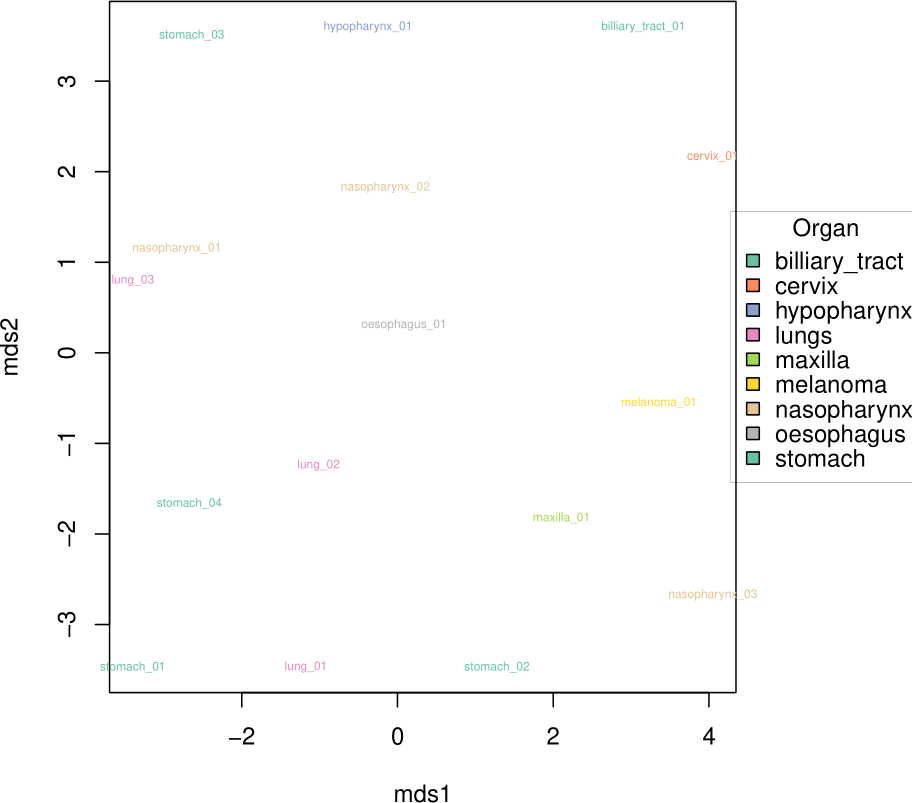


图14 多维尺度（MDS）分析 相同的颜色代表同⼀个组别，彼此间越靠近的样品的相似性越高。

# 关键术语

* TCR: T细胞 (抗原) 受体( T cell receptor ,TCR )
* count 支持克隆的fragments（reads）数量
* freq 克隆频率
* cdr3nt CDR3区核酸序列
* cdr3aa CDR3区氨基酸残基序列
* v V基因
* d D基因
* j J基因
* 克隆多样性指标: 反映每个样品克隆种类的数目和每种克隆频率分布的均一性，包括clone number，shannon diversity，evenness，inverseSimpsonIndex等。

# 参考文献

[1] Shugay M et al. VDJtools: Unifying post-analysis of T cell receptor repertoires. PLoS Comput Biol. 11,11(11)(2015).

[2] Rosati. Overview of methodologies for T-cell receptor repertoire analysis. BMC Biotechnology. (2017)

[3] Hill MO. Diversity and Evenness: A unifying notation and its consequences. Ecology.54,427-432(1973).