2025 年国际定向进化大赛实验记录

实验时间： **2025** 年 **6** 月 **29** 日 **18:00** — **2025** 年 **6** 月 **30** 日 **21:30**

---

23. 全基因组数据分析

一、实验目的

通过全基因组测序（Whole Genome Sequencing, WGS）技术系统筛选人膀胱移行细胞癌 T-24 细胞及其低、中、高浓度 RC48 耐药株（T24-RC48）基因组水平的变异（包括单核苷酸变异、插入/缺失、拷贝数变异及结构变异等），揭示与 RC48 获得性耐药相关的潜在驱动基因及分子机制，为后续功能验证及耐药通路研究提供基因组学依据。

二、实验流程

2.1 样本准备

细胞株：亲本细胞 T24 细胞株及低、中、高耐药 T24-RC48 细胞株（每组 3 个生物学重复）。

DNA 提取：使用高纯度基因组 DNA 提取试剂盒（如 QIAGEN DNeasy Kit），通过琼脂糖凝胶电泳及 Nanodrop 检测 DNA 质量（OD260/280=1.8-2.0，浓度≥50 ng/μL）。

2.2 文库构建与测序

文库制备：采用 Illumina TruSeq DNA PCR-Free Library Prep Kit 构建全基因组文库，片段化至 350 bp，并进行末端修复、接头连接及纯化。

测序平台：Illumina NovaSeq 6000 平台，双端测序（PE150），目标测序深度≥30×。

2.3 数据分析

原始数据质控：FastQC 评估测序数据质量，Trimmomatic 过滤低质量 reads（Q<20，长度<50 bp）。

基因组比对：使用 BWA-MEM 将 clean reads 比对至人类参考基因组（GRCh38/hg38）。

变异检测：

SNV/InDel：GATK HaplotypeCaller 进行变异检测，ANNOVAR 注释变异功能。

拷贝数变异（CNV）：CNVkit 分析拷贝数变化。

结构变异（SV）：Manta 和 Delly 检测染色体结构变异。

耐药相关基因筛选：对比耐药株与亲本株的变异频率，筛选显著富集（$p<0.05$）且与耐药表型相关的候选基因。

2025 年国际定向进化大赛实验记录

实验时间： 2025 年 6 月 29 日 18:00 — 2025 年 6 月 30 日 21:30

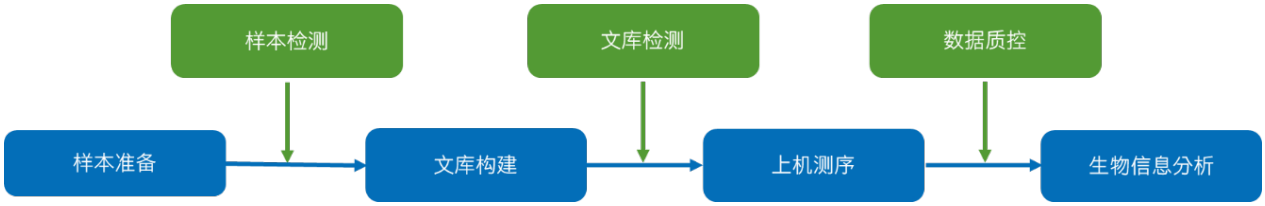

图 1 项目流程图

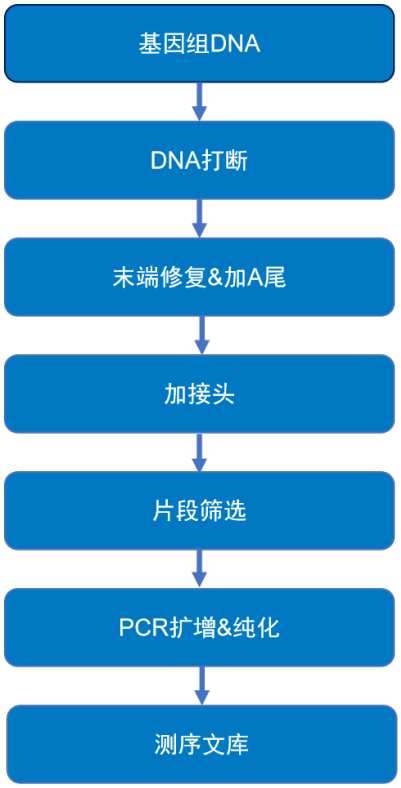

图 2 建库流程

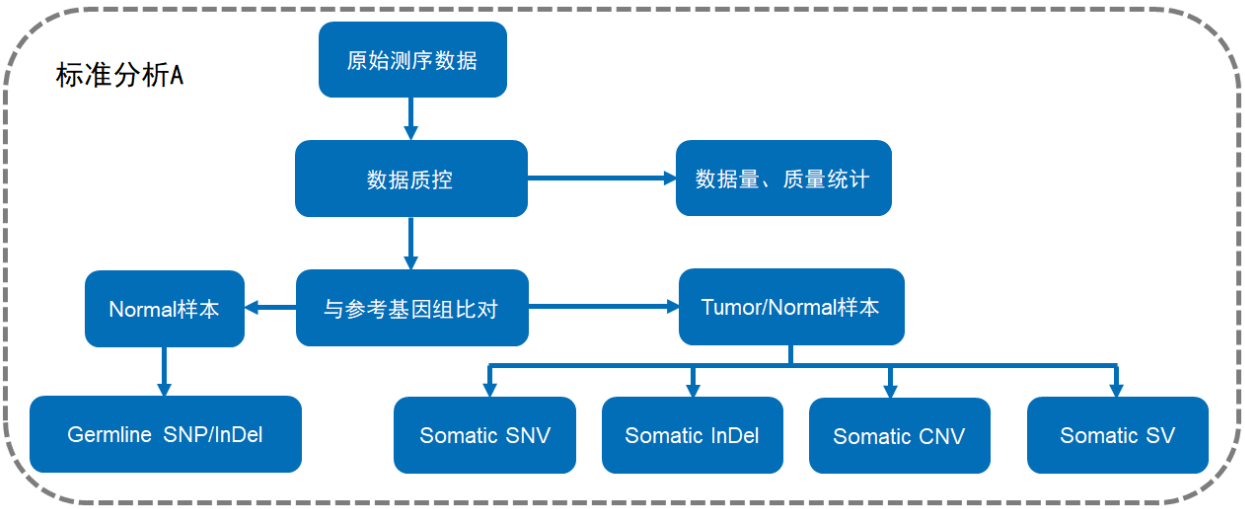

图 3 信息分析流程图

三、实验结果

3.1 聚类分析

2025 年国际定向进化大赛实验记录

实验时间： 2025 年 6 月 29 日 18:00 — 2025 年 6 月 30 日 21:30

利用样本的 SNP 数据进行样本间的聚类分析，以判断正常样本和肿瘤样本是否配对。如图 4 所示：横坐标为各个样本，纵坐标为距离，数值越接近 1 表示样本配对效果越好。

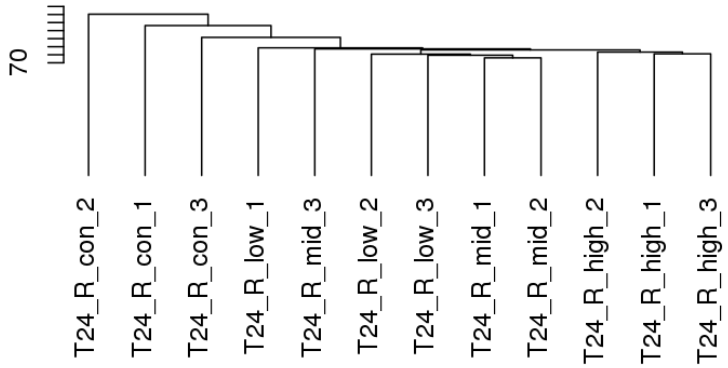

图 4 样本间的聚类分析

3.2 测序数据过滤

测序得到的原始测序序列，里面含有带接头的、低质量的 reads，会对后续信息分析造成很大干扰。为了保证信息分析质量，必须对 raw reads 进行精细过滤，得到 clean reads，后续分析都基于 clean reads。数据处理的步骤如下：

(1) 去除带接头(adapter)的 reads；

(2) 去除 N(N 表示无法确定碱基信息)的比例大于 10%的 reads；

(3) 当单端测序 read 中含有的低质量(低于 5)碱基数超过该条 read 长度比例的 50% 时，需要去除此对 paired reads。

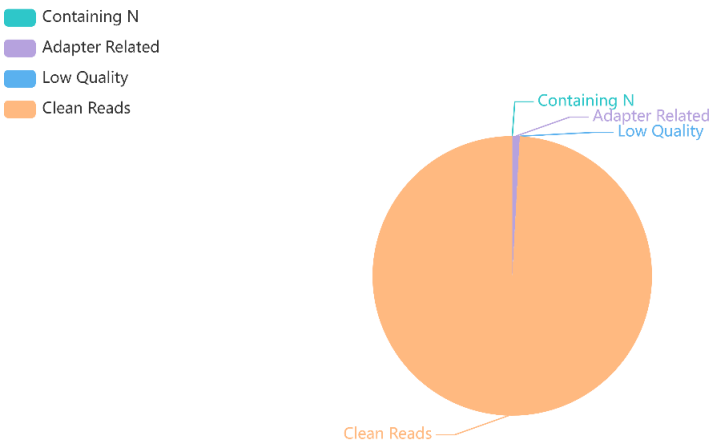

图 5 原始数据过滤结果

3.3 测序错误率分布检查

每个碱基的测序 Phred 值(Phred score, Qphred)是由测序错误率通过公式一转化得到的，而测序错误率是在碱基识别(Base Calling)过程中通过一种判别发生错误概率的模型计算得到

的。对应关系如下表所显示：

表 1 Illumina Casava 1.8 版本碱基识别与 Phred 分值之间的简明对应关系

Phred 分值	不正确的碱基识别	碱基正确识别率	Q-sorce
10	1/10	90%	Q10
20	1/100	99%	Q20
30	1/1000	99.9%	Q30
40	1/10000	99.99%	Q40

测序错误率分布检查用于检测在测序长度范围内，有无某些位置的碱基存在异常的高错误率，例如如果中间位置的碱基测序错误率显著高于其他位置，则可能存在异常碱基。测序错误率与碱基质量有关，受测序仪本身、试剂、样品等多个因素共同影响。对于 Illumina 高通量测序平台，测序错误率分布具有两个特点：

（1）测序错误率会随着测序的进行而升高，这是由于测序过程中荧光标记的不完全切割等因素引起荧光信号衰减，因而导致错误率升高。

（2）每个 Read 前几个碱基的位置也会有较高的测序错误率，这是由于边合成边测序过程初始阶段，测序仪荧光感光元件对焦速度较慢，获取的荧光图像质量较低，导致碱基识别错误率较高。

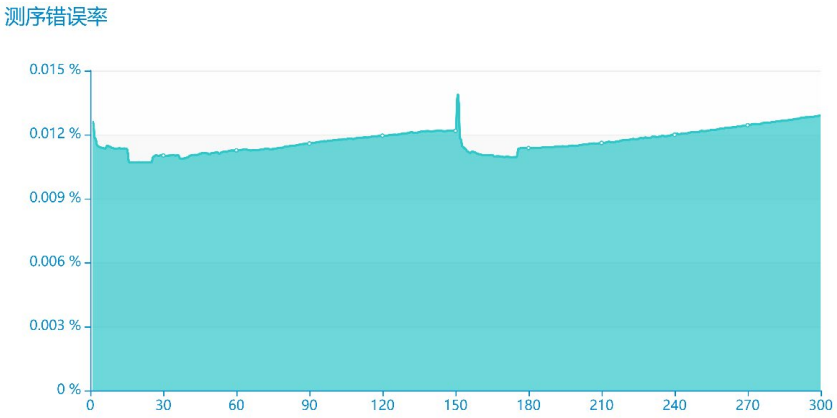

图 6 测序错误率分布图

横坐标为 Reads 的碱基位置，纵坐标为所有 Reads 该位置碱基的平均错误率，左侧 1~150 bp 为 Read1 错误率分布，右侧 151~300 bp 为 Read2 错误率分布。

3.4 测序数据质量分布

测序数据的质量主要分布在 Q30（≥85%）以上，这样能够保证后续分析的正常进行。据测序技术的特点，测序片段末端的碱基质量一般会比前端的低。

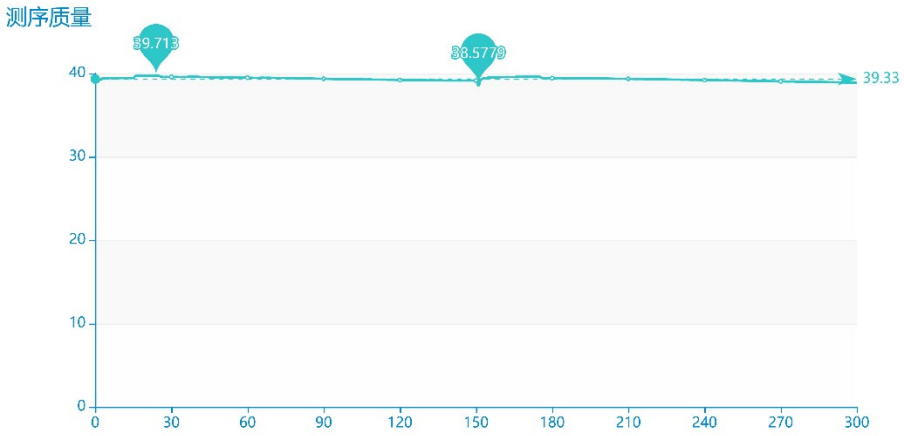

图 7 数据质量分布

横坐标为 Reads 的碱基位置，纵坐标为所有 Reads 该位置碱基的平均质量值（Phred 分值），左侧 1~150 bp 为 Read1 数据质量分布，右侧 151~300 bp 为 Read2 数据质量分布

根据 Illumina NovaSeq 平台的测序特点，使用双端测序的数据，我们要求 Q30 平均比例在 85%以上，平均 Error rate 在 0.1%以下。本次测序样本的平均 Rawdata 为 100.84G，平均有效数据量占比 99.30%，平均 Q20 为 98.80%，平均 Q30 为 96.53%，平均错误率为 0.01%。综上所述本次测序数据质量良好，满足分析需求。

表 2 数据产出质量情况一览表

Sample name	ID	Flowcell/Lane	Raw reads	Raw data(G)	Effective(%)	Error(%)	Q20(%)	Q30(%)	GC(%)
T24_R_mid_1	SKDO250000421-1A	A22M7WYLT4-new_L4	362,928,424	108.88	99.28	0.01	98.93	96.88	40.70
T24_R_high_1	SKDO250000424-1A	A22M7WYLT4-new_L3	348,054,195	104.42	99.32	0.01	98.95	96.97	41.97
T24_R_con_3	SKDO250000417-1A	A22M7WYLT4-new_L1	324,249,414	97.27	99.17	0.01	98.93	96.90	41.24
T24_R_con_2	SKDO250000416-1A	A22M2NNLT4-new_L1	270,230,632	90.34	99.38	0.01	98.61	95.99	41.61
T24_R_con_2	SKDO250000416-1A	A22M7TNLT4-new_L4	30,889,541		99.39	0.01	98.36	95.17	41.61
T24_R_high_3	SKDO250000426-1A	A22M2NNLT4-new_L2	381,505,540	114.45	99.25	0.01	98.47	95.60	41.30
T24_R_low_1	SKDO250000418-1A	A22LF7FLT4-new_L4	82,288,765	90.66	99.14	0.01	98.72	96.44	40.71
T24_R_low_1	SKDO250000418-1A	A22M7WYLT4-new_L1	219,921,244		99.14	0.01	98.87	96.73	40.59
T24_R_low_2	SKDO250000419-1A	A22M7WYLT4-new_L1	333,015,710	99.90	99.33	0.01	98.92	96.91	41.45
T24_R_mid_2	SKDO250000422-1A	A22M7WYLT4-new_L4	375,892,898	112.77	99.39	0.01	98.94	96.91	41.23
T24_R_mid_3	SKDO250000423-1A	A22M7WYLT4-new_L3	305,432,411	91.63	99.42	0.01	98.91	96.88	41.83
T24_R_con_1	SKDO250000415-1A	A22M7TNLT4-new_L4	105,294,094	96.95	99.37	0.01	98.31	95.03	41.29
T24_R_con_1	SKDO250000415-1A	A22M2NNLT4-new_L5	217,859,874		99.36	0.01	98.32	95.00	41.19
T24_R_low_3	SKDO250000420-1A	A22M7WYLT4-new_L4	341,431,741	102.43	99.19	0.01	98.91	96.84	40.84
T24_R_high_2	SKDO250000425-1A	A22M7WYLT4-new_L3	334,716,355	100.41	99.41	0.01	98.94	96.95	41.72

3.5 测序深度、覆盖度分布

有效测序数据通过 BWA 比对到参考基因组，得到 BAM 格式的最初的比对结果。然后，用 Sambamba 对比对结果进行排序；并标记重复 reads（mark duplicate reads）。最后，我们利用重复标记后的比对结果进行覆盖度、深度等的统计。通常，人类样本的测序 reads

能达到 95%以上的比对率；当一个位点的碱基覆盖深度（read depth）达到 10X 以上时，该位点处检测出的 SNV 比较可信。

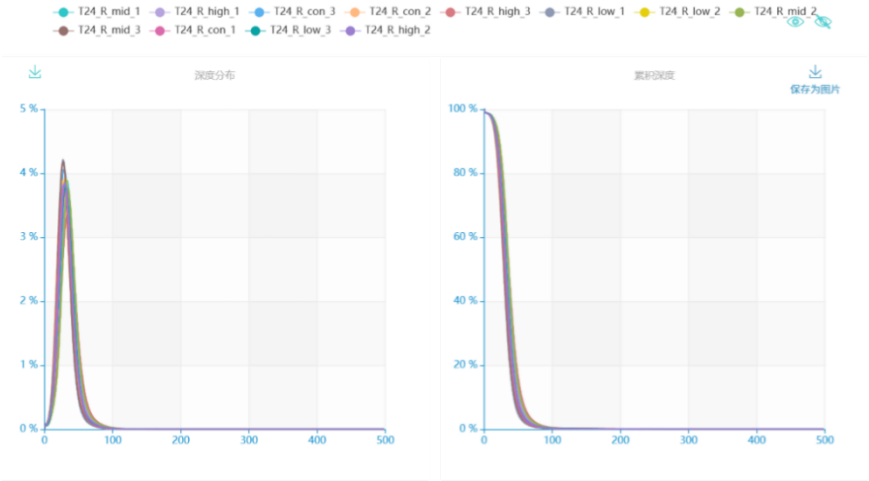

图 8 测序深度

左图为不同测序深度的碱基比例，横坐标表示测序深度，纵坐标表示测序深度为 x 的碱基在所有碱基中的比例；图像一般在平均深度周围成泊松分布；右图为不同深度上的累积碱基比例，横坐标表示测序深度，纵坐标代表测序深度超过 x 的碱基在所有碱基中的比例，比如测序深度为 0 对应了碱基比例 100%，表示有 100%的碱基其测序深度大于 0。

图 9 每个染色体的覆盖深度 (左侧坐标) 和覆盖率 (右侧坐标)

横坐标表示染色体编号，左侧纵坐标表示平均覆盖深度，右侧纵坐标表示覆盖率。对每条染色体计算覆盖深度时，计算公式为：每条染色体的测序数据量/每条染色体上外显子区域的总长度。计算覆盖率时，公式为：每条染色体被覆盖的总长度/每条染色体上外显子区域的总长度。

3.6 覆盖度统计结果

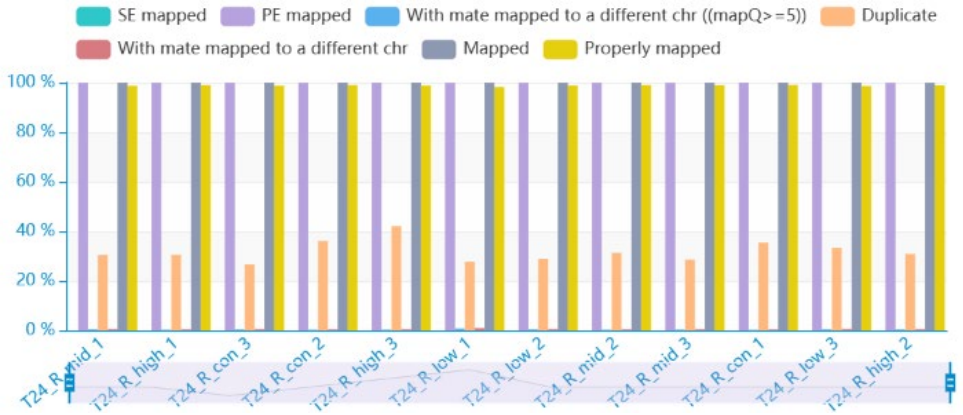

图 10 各特性 reads 占总 reads 数量的比值

3.7 SNP 统计结果

通常，一个人全基因组内会有约 3.6M 个 SNP，绝大数 (大于 95%) 的高频 (群体中等位基因频率大于 5%) 的 SNP 在 dbSNP(Sherry et al., 2001)中有记录。转换/颠换的比值 (Ts:Tv) 可以反应 SNP 数据集的准确性，全基因组内的比值约在 2.2 左右，编码区内的比值约在 3.2 左右。SNP 的检测和统计结果展示如下：

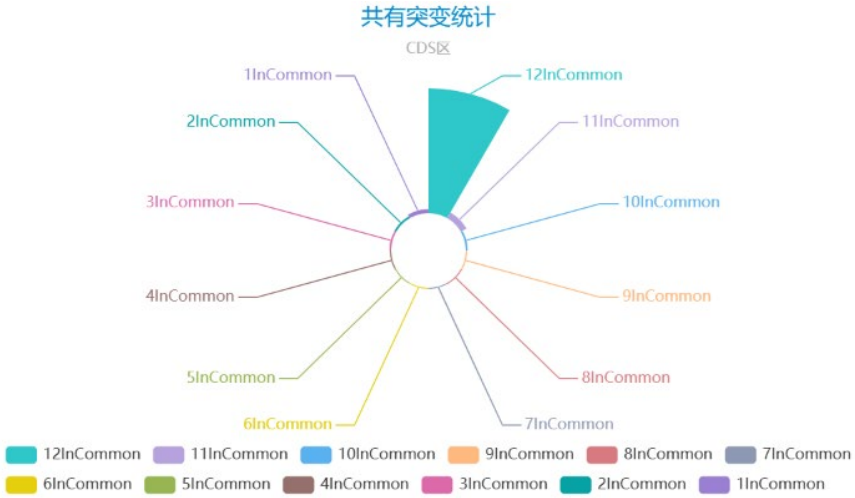

图 11 CDS 区共有 SNP 突变统计（3InCommon 表示在 3 个样本中共有的突变数）

在比对结果的基础上，我们利用 bcftools(Li et al., 2009)识别 SNP 位点，并采用国际惯用的过滤标准对 SNP 位点进行过滤。本次测序样品中，平均每个样本发现 3155367 个 SNP，其中 19802 出现在外显子区，1718276 个在基因间区，546 个在 splicing 区。在编码区的 SNP 中，平均每个样本有 10017 个同义突变，9245 个错义突变，69 个 SNP 导致该碱基所在的密码子变为终止密码子，有 10 个 SNP 导致该碱基所在的终止密码子变为非终止密码子。SNP 统计结果具体如下表所示：

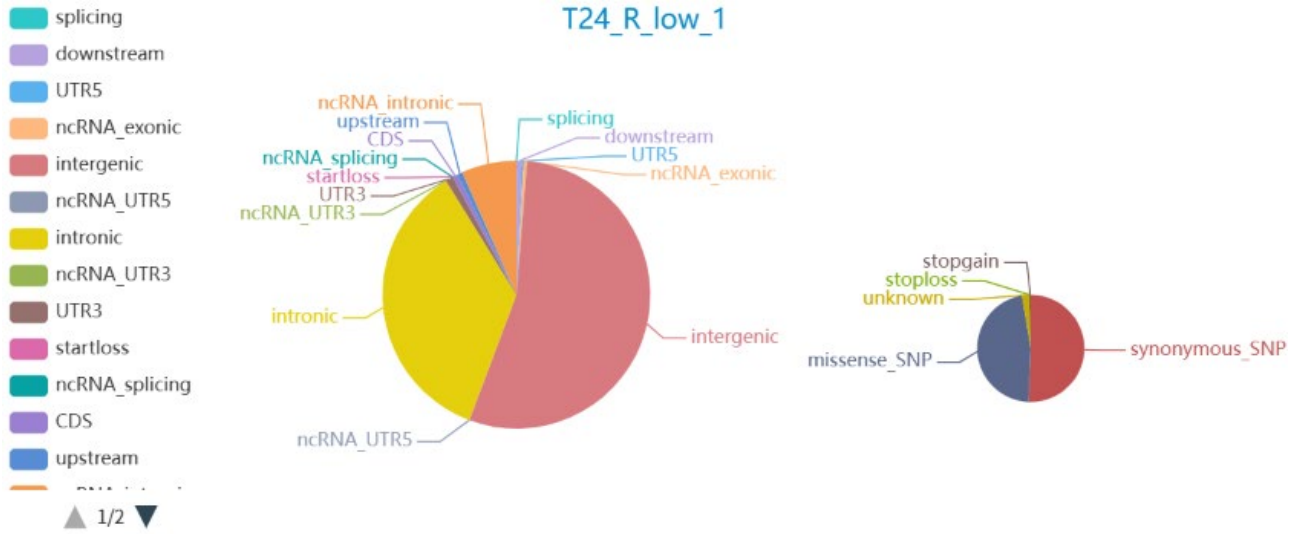

图 12 基因组不同区域上 SNP 数目 (左) 和编码区上不同类型的 SNP 数目 (右) 分布

图 13 基因组 SNP 特征

3.8 InDel 统计结果

通常，一个人全基因组内会有约 350K 的 InDel (insertion and deletion，小于 50bp 的插入缺失) 。编码区或剪接位点处发生的插入缺失都可能会改变蛋白的翻译。移码变异，其插入或缺失的碱基串的长度为非 3 的整数倍，因此可能导致整个读框的改变；与非移码变异比较，受到的限制更大。

在比对结果的基础上，我们利用 bcftools(Li et al., 2009)识别 InDel，并采用国际惯用的过滤标准对 InDel 结果进行过滤。本次测序样品中，平均每个样本发现 923236 个 InDel，其中 571 出现在外显子区，468933 个在基因间区，385 个在 splicing 区。在编码区的 InDel 中，平

均每个样本有 68 个移码缺失，61 个移码插入，177 个非移 码缺失，149 个非移码插入，3 个 InDel 导致丢失终止密码子，10 个 InDel 导致获得终止密码子。InDel 结果统计如下：

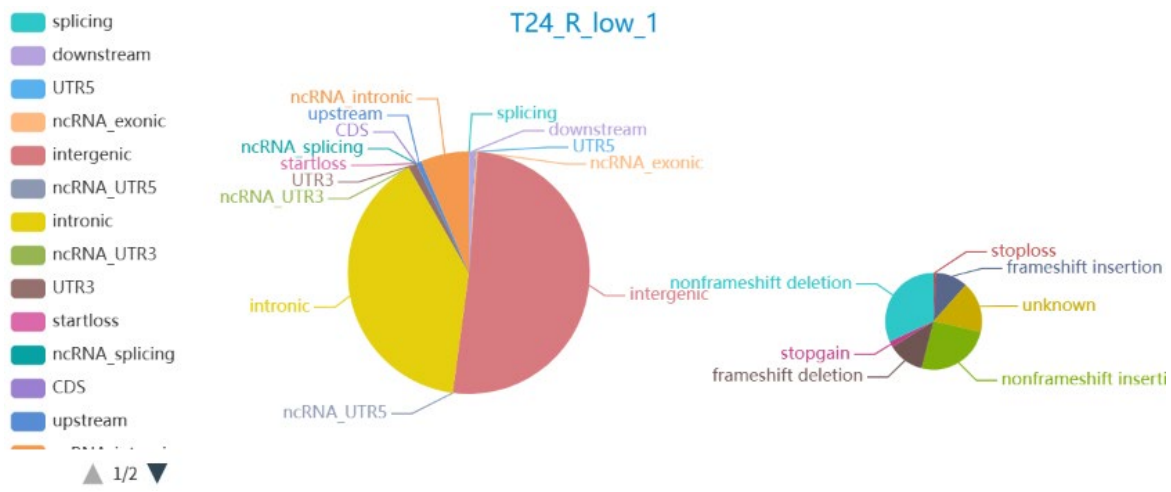

图 14 基因组不同区域上 InDel 数目(左) 和编码区上不同类型的 InDel 数目 (右) 分布

图 15 基因组 InDel 特征

3.9 体细胞变异检测(PON)

体细胞突变 (Somatic mutation) 是指除生殖细胞以外的体细胞所发生的突变，是发生在正常机体细胞中的突变，如发生在皮肤和器官。体细胞突变既不遗传自亲本，也不会传递给后代，却可以引起当代某些细胞的遗传结构发生改变。体细胞突变，特别是其中的驱动突变 (Driver mutation) 对解释肿瘤的发生和发展具有非常重要的意义，另一方面肿瘤耐药性的产生也与体细胞突变有关。因此，关注体细胞突变是肿瘤基因组研究的重心，也是肿瘤基因组研究区别于疾病研究的一个特性。

SNV 全称 Single nucleotide variant，是指在基因组上由单个核苷酸的替换所引起的变异。我们主要使用 muTect2 软件的 PON 模式(https://gatk.broadinstitute.org/hc/en-

us/articles/360036364532-CreateSomaticPanelOfNormals-BETA-)来寻找 Somatic SNV 位点。

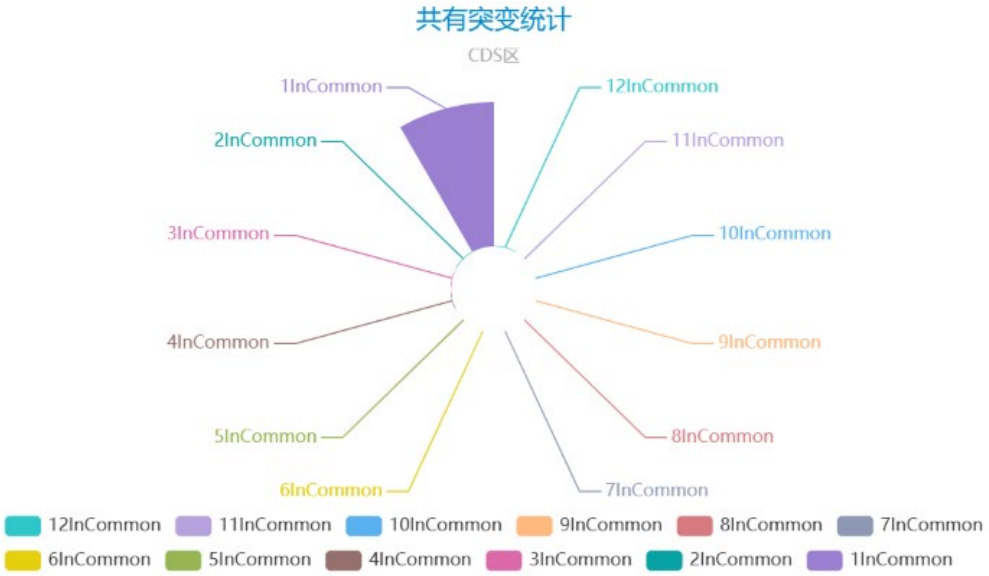

图 16 CDS 区共有 Somatic SNV 突变统计（3InCommon 表示在 3 个样本中都存在的突变）

四、结果分析

耐药细胞株中筛选出的高频变异基因及基因组结构变化，提示其可能在 RC48 耐药性的获得中发挥潜在调控作用。通过对比亲本株与不同耐药程度细胞株的变异谱，可观察到耐药相关基因的逐步富集趋势，表明基因组动态演化与耐药表型之间存在显著相关性。

关键基因的功能性突变（如激酶域或蛋白功能域变异）可能通过影响药物靶点的结合效率或下游信号转导能力，直接或间接导致耐药性增强。此外，拷贝数变异（如特定基因的扩增或缺失）可能通过改变关键信号通路的活性，促进细胞存活或凋亡逃逸，进而形成耐药表型。

耐药细胞株中检测到的结构变异及大片段重排事件，可能通过驱动基因组不稳定性加速耐药相关突变的累积。此类变异可能影响细胞周期调控、DNA 修复或表观遗传修饰相关基因的功能，从而为耐药克隆的适应性进化提供遗传基础。

全基因组变异数据提示，耐药性的形成可能涉及多基因、多通路的协同作用。例如，靶点基因的修饰突变与促存活信号通路的激活可能共同削弱药物疗效，而 DNA 修复相关基因的异常可能进一步加剧耐药克隆的筛选压力耐受性。

基因组学数据的生物学意义需结合转录组、蛋白质组及功能实验进行整合验证，以明确候选基因在耐药中的具体作用。此外，耐药相关变异的进化规律及其与临床耐药的相关性仍需在更复杂的模型或临床样本中进一步探索。

综上所述，全基因组测序从系统层面揭示了耐药细胞株的基因组变异特征，为解析

2025 年国际定向进化大赛实验记录

实验时间： **2025** 年 **6** 月 **29** 日 **18:00** — **2025** 年 **6** 月 **30** 日 **21:30**

RC48 获得性耐药机制提供了重要的候选基因和分子线索。后续研究需聚焦于关键变异的生物学功能验证及多组学数据的联合分析，以全面阐明耐药机制的网络调控模式。