

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T24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获得耐药相关和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($OD_{260}/OD_{280} = 1.8 - 2.0$, 浓度 $\geq 50 \mu g/\mu l$)。

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300

2.3 数据分析

原始数据分析:FastQC评估测序数据质量,Trimmmate过滤低质量reads (QC20, 长度<50 bp)。

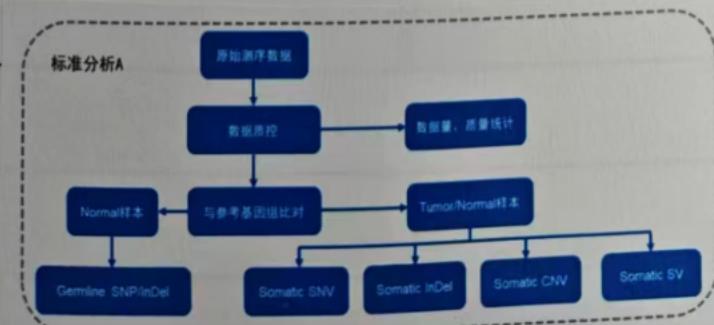
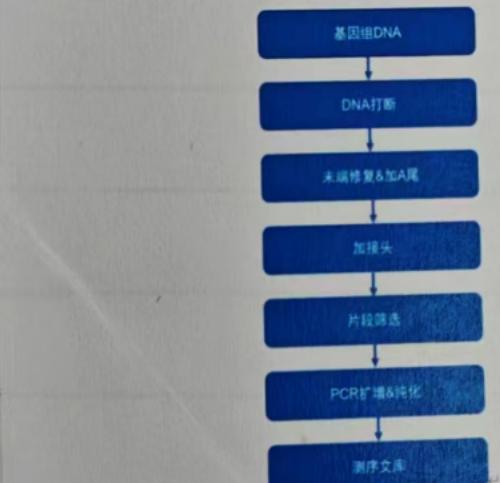
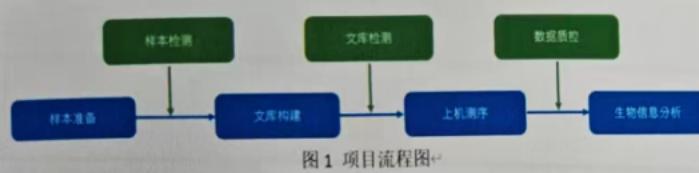
基因组比对: 使用 BWA MEM 将 clean reads 对比至人类参考基因组 (GRCh38/hg38)。

变异检测: SNV/Indel: GATK Haplotype Caller 进行变异检测, ANNOVAR 注释。

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。

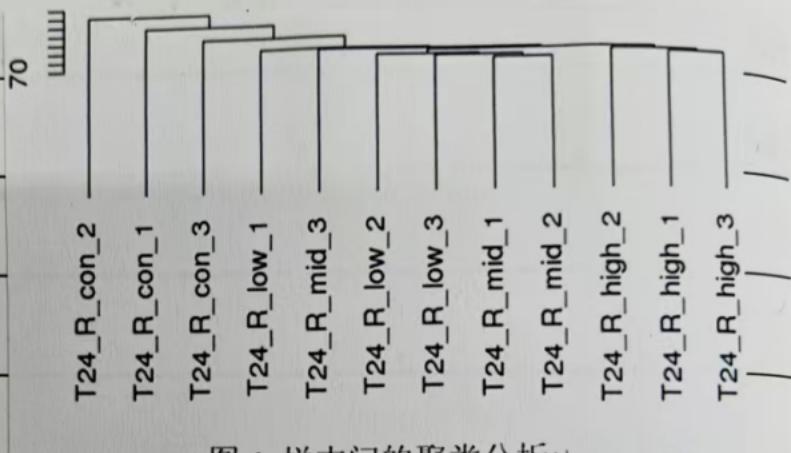
耐药相关基因筛选: 对比耐药株与亲本株的变异频率, 筛选显著富集 ($p < 0.05$) 且与耐药表型相关的候选基因。



三、实验结果

3.1 聚类分析

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，数值越接近表示样本配对越好。



3.2 测序数据过滤

测序得到的原始数据测序序列，里面有带有接头的、低质量 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。数据处理步骤如下：

(1) 去除带头(adapter) reads；

(2) 去除 N (N 表示无法确定碱基信息) reads

比例大于 10% in reads；

(3) 当单端测序 read 中含有低质量(低

于 5) 碱基数超过该条 read 长度比例

在 50% 时需去除此对 paired reads.

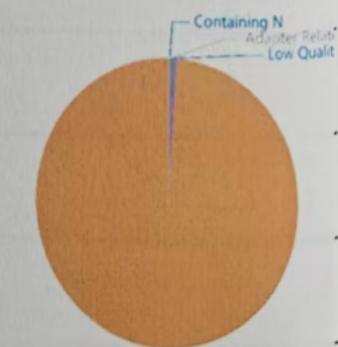
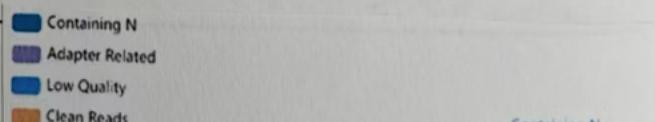


图 5 原始数据过滤结果

3.3 测序信息错误率分布检查

每个碱基的测序 Phred 值 (Phred score, Q_{phred}) 是由测序错误率通过公式转化得到的，而测序错误率是在碱基识别过程中通过一种判别发生错误概率的模型计算得来的。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-score
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。测序错误率与碱基质量有关，受测序仪本身、试剂、样品等众多因素共同影响。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，导致碱基识别率较高。



图 6 测序错误率分布图

横坐标为 Reads 在碱基位置，纵坐标为所有 Reads 该位置碱基平均质量值 (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 平台测序特点，使用双端测序的数据，我们要求 Q30 平均值要在 85% 以上，平均 Error rate 在 0.1% 以下。本次测序样本的平均 Raw data 为 100.8G，平均有效数据量占比 99.30%，平均 Q20 为 98.8%，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。

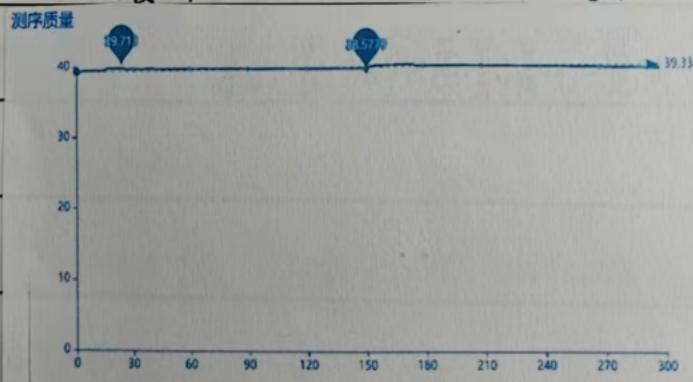


图 7 数据质量分布

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	SAD0250000021_1A	A22M7WY1T4_new_L4	362,504,424	108.88	99.26	0.02	96.91	96.89	42.70
T24_R_high_1	SAD0250000042_1A	A22M7WY1T4_new_L3	348,264,195	104.40	99.32	0.02	96.95	96.97	41.87
T24_R_low_3	SAD0250000017_1A	A22M7WY1T4_new_L1	324,249,614	97.27	99.17	0.01	96.97	96.90	41.24
T24_R_low_2	SAD0250000041_1A	A22M7WY1T4_new_L2	270,230,832	90.34	99.38	0.01	96.62	95.99	41.61
T24_R_low_2	SAD0250000016_1A	A22M7WY1T4_new_L6	30,889,542	—	99.29	0.01	96.36	95.17	41.61
T24_R_low_3	SAD0250000046_1A	A22M7WY1T4_new_L2	381,505,540	114.49	99.25	0.01	96.47	95.62	41.30
T24_R_low_1	SAD0250000018_1A	A22L7P1T14_new_L4	82,286,705	96.46	99.14	0.01	96.72	96.44	40.71
T24_R_low_1	SAD0250000018_1A	A22M7WY1T4_new_L2	229,921,244	—	99.14	0.01	96.87	96.73	40.59
T24_R_low_2	SAD0250000041_1A	A22M7WY1T4_new_L5	333,015,715	99.90	99.23	0.02	96.92	96.91	42.45
T24_R_mid_2	SAD0250000042_1A	A22M7WY1T4_new_L4	375,092,096	112.77	99.39	0.02	96.94	96.95	41.23
T24_R_low_3	SAD0250000042_1A	A22M7WY1T4_new_L3	305,412,611	91.13	99.42	0.02	96.31	96.38	41.83
T24_R_low_1	SAD0250000045_1A	A22M7WY1T4_new_L4	358,294,994	96.95	99.27	0.02	96.33	96.38	41.29
T24_R_low_3	SAD0250000045_1A	A22M7WY1T4_new_L5	237,359,374	95.36	99.36	0.02	96.32	95.98	41.29

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BA从格式的最初比对结果。然后，用Samtools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达到 10X 以上时，该位点处检测出 SNV 较可信。

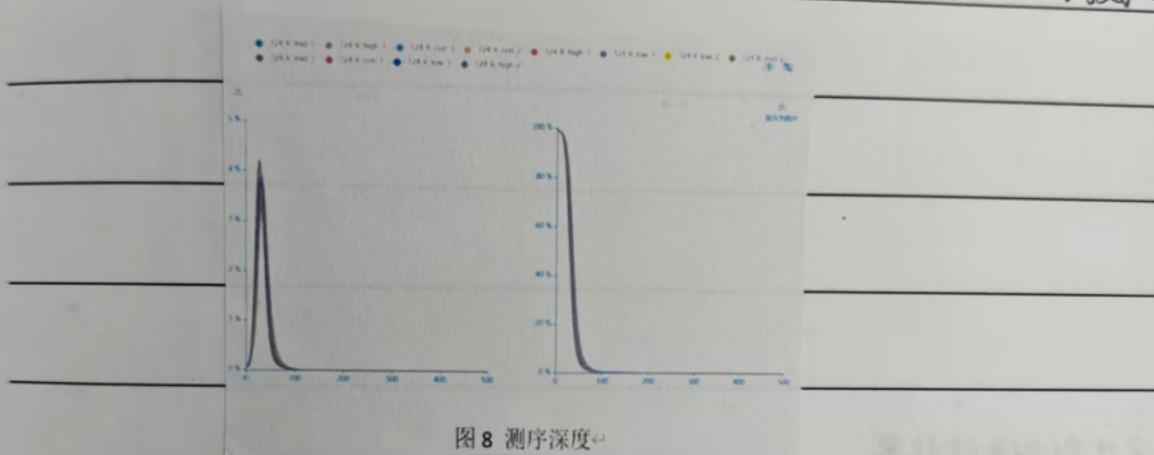


图 8 测序深度

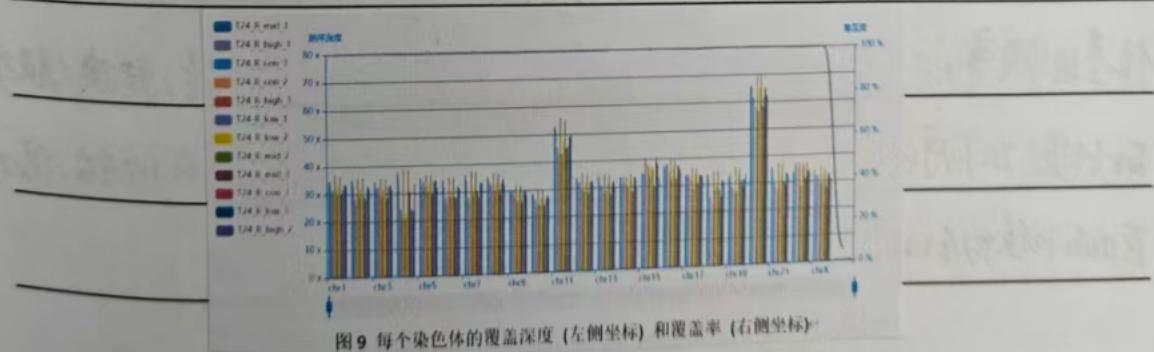


图 9 每个染色体的覆盖深度 (左侧坐标) 和覆盖率 (右侧坐标)

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，一个全基因组内会有约3.6M个SNP，绝大多数（大于90%）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，编码区内约在3.2左右。SNP的检测和统计结果展示如下：



在对比结果的基础上,我们利用bcftools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3135387个SNP,其中19802出现在外显子区,1718276个在基因间区,146个在splicing区。在编码区的SNP中平均每个样本有1001个同义突变,9245个错义突变,69个SNP导致读碱基所在的位置码变为终止密码子,有10个SNP导致读碱基所在的位置码变为非终止密码子。

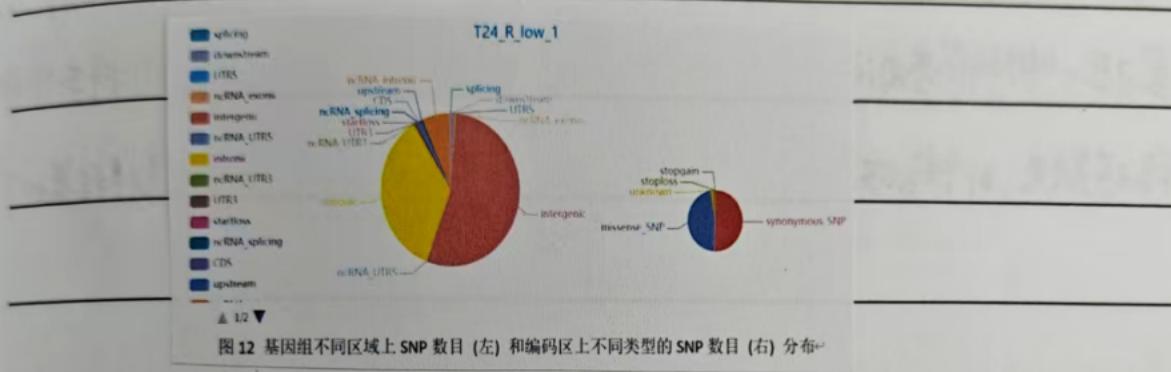


图 12 基因组不同区域上 SNP 数目 (左) 和编码区上不同类型的 SNP 数目 (右) 分布

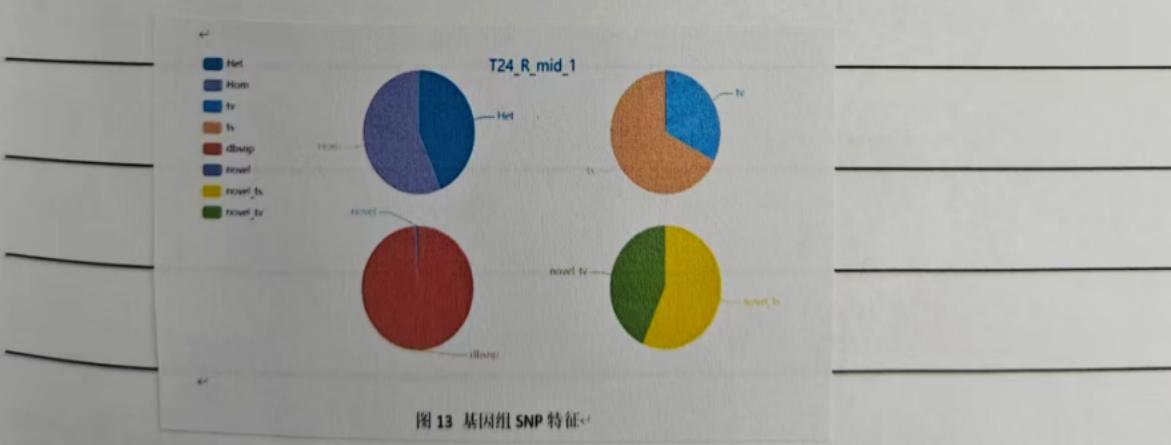


图 13 基因组 SNP 特征

3.8 InDel 统计结果

通常一个全基因组内会有约350K个 InDel (insertion and deletion, 小于50bp的插入缺失). 编码区或剪接位点发生的插入缺失都可能会改变蛋白的翻译。移码突变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.

在此对结果的基础上, 我们利用比对识别 InDel, 并采用国际通用的过滤标准对 InDel 结果进行过滤。本次测序平均每个样本发现 923236 个 InDel, 其中 571 个出现在外显子区, 468933 在基因间区, 355 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。统计结果如下:

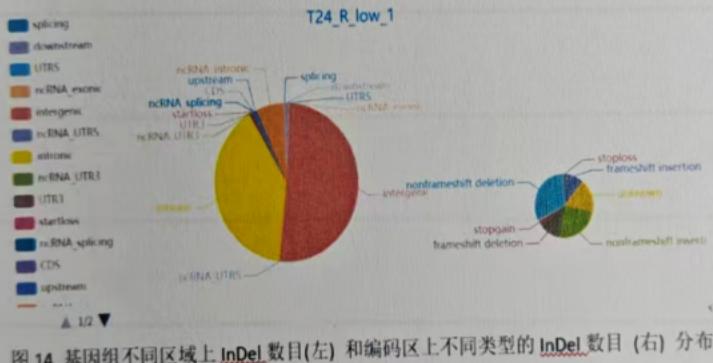


图 14 基因组不同区域上 InDel 数目(左) 和编码区上不同类型的 InDel 数目(右) 分布

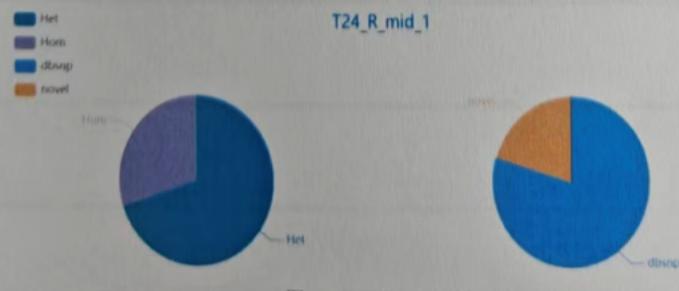


图 15 基因组 InDel 特征

3.9 体细胞变异检测 (SOMV)

体细胞突变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。因此，关注体细胞突变是肿瘤基因组研究的重心，也是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。主要使用 muTect2 软件的 PON 模式 (<http://gatk.com>) 来寻找 Somatic SNV 位点。

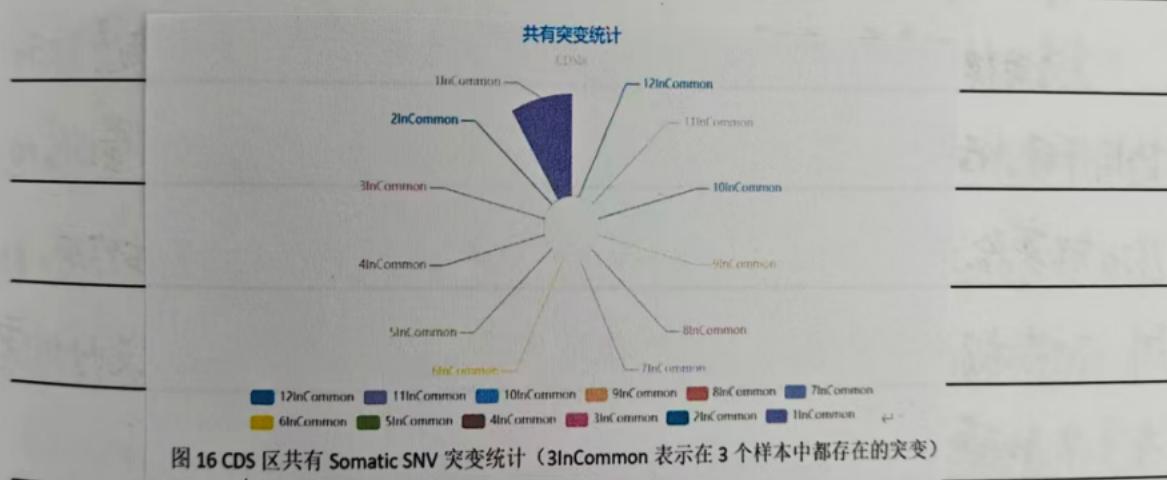


图 16 CDS 区共有 Somatic SNV 突变统计 (3InCommon 表示在 3 个样本中都存在的突变)

四 结果分析

耐药细胞株中筛选出的高频率突变基因及基因组结构变化，提示其可能在 KC48 耐药性形成过程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。

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修复或表现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。而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复杂的模型或临床样本中进一步探索。