International Directed Evolution Competition Lab Notebook

Experiment time: 2025-06-29, 18:00 - 2025-06-30, 21:30

Whole Genome Data Analysis

I. Objective of the experiment

Through whole genome sequencing (WGS) technology, we systematically screened genomic variations (including single nucleotide variants, insertions/deletions, copy number variations, and structural variations) in human bladder transitional cell carcinoma T-24 cells and their low-, medium-, and high-grade RC48-resistant sublines (T24-RC48). This study identified potential driver genes and molecular mechanisms associated with RC48-induced drug resistance, providing genomic evidence for subsequent functional validation and drug resistance pathway research.

II. Experimental procedures

2.1 Sample preparation

Cell lines: Parental cell T24 cell line and low, medium and high drug resistant T24-RC48 cell line (3 biological repeats per group).

DNA extraction: Use a high-purity genomic DNA extraction kit (e.g., QIAGEN DNeasy Kit), and detect the DNA quality by agarose gel electrophoresis and Nanodrop (OD260/280=1.8-2.0, concentration ≥50 ng/μL).

2.2 Library construction and sequencing

Library preparation: Illumina TruSeq DNA PCR-Free Library Prep Kit was used to construct the whole genome library, which was fragmented to 350 bp, and then end repair, linker connection and purification were performed.

Sequencing platform: Illumina NovaSeq 6000 platform, dual-end sequencing (PE150), target sequencing depth ≥30×.

2.3 Data Analysis

Quality control of raw data: FastQC was used to evaluate the quality of sequencing data, and Trimmomatic was used to filter out low-quality reads (Q score < 20, length < 50 bp).

Genome alignment: Clean reads were aligned to the human reference genome (GRCh38/hg38) using BWA-MEM.

Mutation testing:

SNV/InDel: GATK HaplotypeCaller is used to detect variations, and ANNOVAR was used

for functional annotation of the identified variants.

Copy number variation (CNV): CNVkit analyzes copy number variations.

Structural variation (SV): Manta and Delly detect chromosomal structural variation.

Drug resistance-related gene screening: The mutation frequencies in resistant cell lines were compared with the parental line to screen for candidate genes that were significantly enriched ($p < 0.05$) and associated with the drug-resistant phenotype.

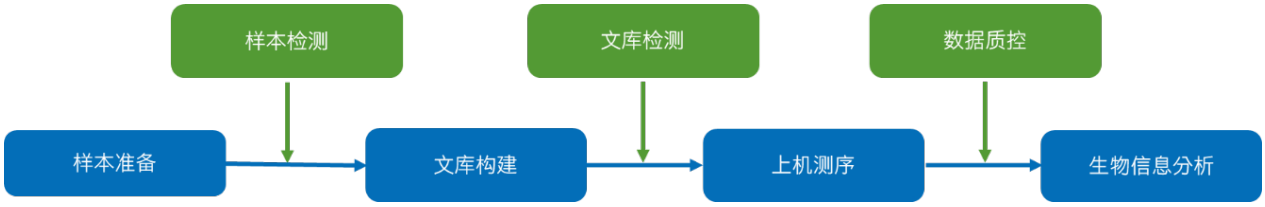

Figure 1 Project flow chart

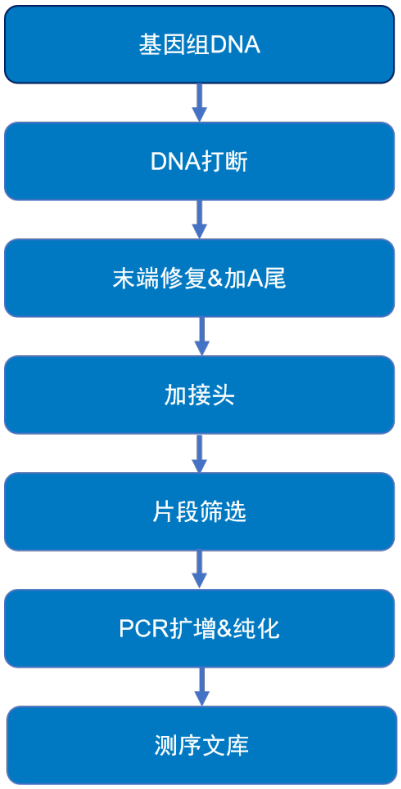

Figure 2 Library building process

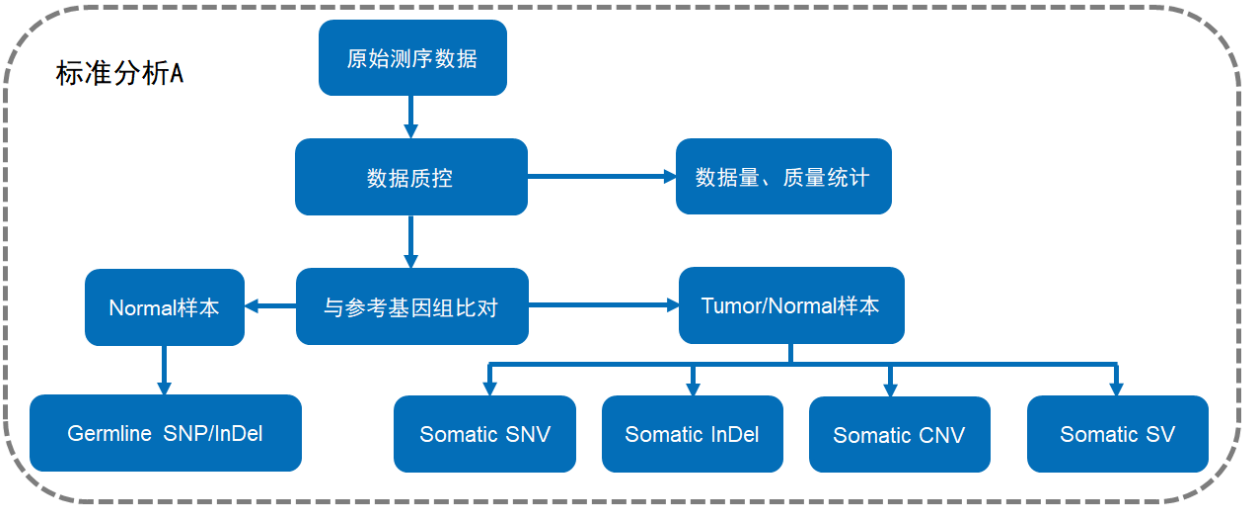

Figure 3 Information analysis flow chart

III. Experimental results

3.1 Cluster analysis

The SNP data of the samples were used for inter-sample clustering analysis to assess the genetic relatedness and clustering patterns among the parental and drug-resistant cell lines. As shown in Figure 4: the horizontal coordinate is each sample, and the vertical coordinate is the distance, and the closer the value is to 1, the better the pairing effect of the samples.

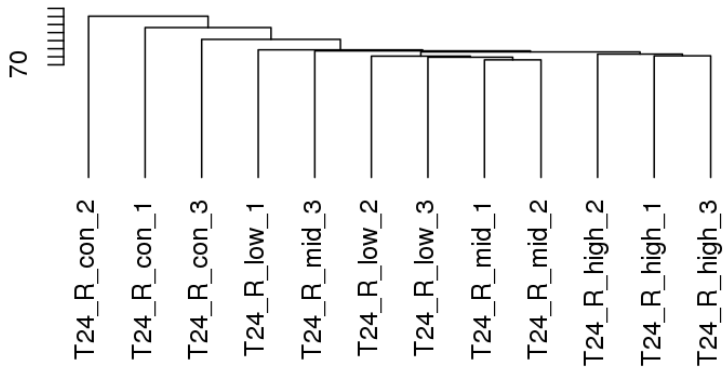

Figure 4 Cluster analysis between samples

3.2 Sequencing data filtering

The raw sequencing sequences obtained through sequencing contain low-quality, adapter-ligated reads that can significantly interfere with subsequent data analysis. To ensure the quality of information analysis, it is essential to perform rigorous filtering on raw reads to obtain clean reads, which will serve as the foundation for all subsequent analyses. The data processing steps are as

follows:

(1) Remove reads with adapters;

(2) Remove reads with a proportion of 'N' bases (where 'N' indicates an undetermined base) greater than 10%;

(3) When the number of low quality (less than 5) base numbers in a single end sequencing read exceeds 50% of the length of the read, the paired read pair should be removed.

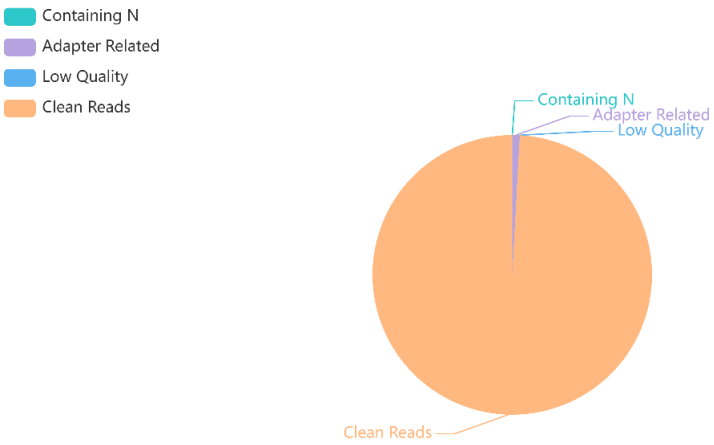

Figure 5 Original data filtering results

3.3 Distribution of sequencing error rate

The Phred score (Qphred) of each base is obtained by transforming the sequencing error rate through formula 1, and the sequencing error rate is calculated by a model that determines the probability of error during the base calling process. The corresponding relationship is shown in the following table:

Table 1 A concise correspondence between base identification and Phred scores in version 1.8 of Illumina Casava

Phred value	**Incorrect base recognition**	**Correct identification rate of base**	**Q-sorce**
10	1/10	90%	Q10
20	1/100	99%	Q20
30	1/1000	99.9%	Q30
40	1/10000	99.99%	Q40

The sequencing error rate distribution analysis identifies abnormal high-error-rate positions within the sequencing length range. For instance, if base errors are significantly higher in mid-sequence regions compared to other positions, this may indicate abnormal base occurrences. These error rates correlate with base quality, influenced by multiple factors including sequencing instruments, reagents, and sample conditions. For Illumina high-throughput sequencing platforms, the error rate distribution exhibits two distinct characteristics:

(1) The sequencing error rate increases with the progress of sequencing. This is because the fluorescence signal attenuation caused by incomplete cutting of fluorescent markers during sequencing leads to the increase of error rate.

(2) The position of the first few bases of each Read also has a high sequencing error rate, which is due to the slow focusing speed of the fluorescent photosensitive element of the sequencer in the initial stage of the process of simultaneous synthesis and sequencing, and the low quality of the obtained fluorescence image, leading to a high base identification error rate.

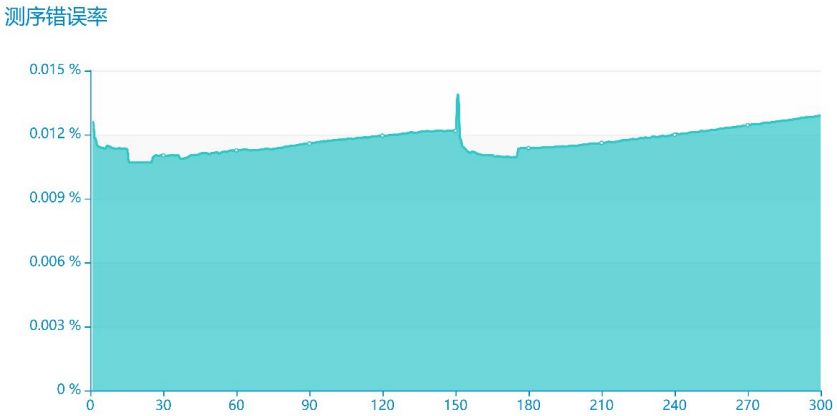

Figure 6 Distribution of sequencing error rates

The horizontal axis is the base position of Read, and the vertical axis is the average error rate of all Read bases at that position. Positions 1-150 bp (left) represent the error rate distribution for Read1, and positions 151-300 bp (right) represent that for Read2.

3.4 Distribution of sequencing data quality

The quality of the sequencing data is mainly distributed above Q30 (≥85%), which can ensure the normal progress of subsequent analysis. According to the characteristics of sequencing technology, the base quality at the end of the sequencing fragment is generally lower than that at the

front.

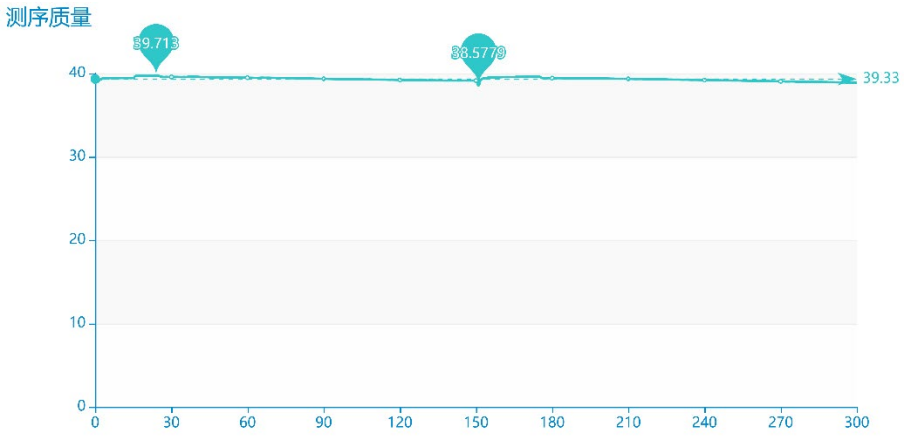

Figure 7 Data quality distribution

The horizontal axis represents the base position of Reads, and the vertical axis represents the average quality value (Phred score) of all bases at that position in Reads. The left 1-150 bp shows the data quality distribution of Read1, and the right 151-300 bp shows the data quality distribution of Read2

Based on the sequencing characteristics of Illumina NovaSeq platform, when using dual-end sequencing data, we require an average Q30 ratio above 85% and an average error rate below 0.1%. The average Rawdata of this sequencing sample was 100.84G, with 99.30% of data being valid. The average Q20 reached 98.80%, while the average Q30 stood at 96.53%, achieving an average error rate of 0.01%. In conclusion, the sequencing data quality meets the analysis requirements and demonstrates satisfactory performance.

Table 2 Overview of data output quality

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 Distribution of sequencing depth and coverage

Effective sequencing data undergoes BWA alignment against the reference genome to obtain initial alignment results in BAM format. The alignment results are then sorted using Sambamba and duplicates are marked. Finally, statistical analysis of coverage depth and other metrics is performed on the marked alignment results. Typically, human samples achieve over 95% read alignment rate. When a site's base coverage depth (read depth) exceeds 10X, the detected SNVs at that site are considered reliable.

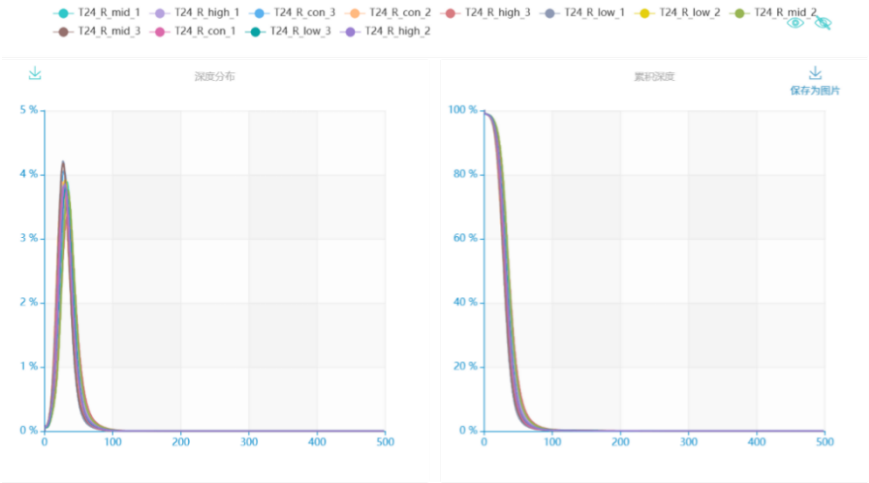

Figure 8 Sequencing depth

The left figure illustrates base proportions at different sequencing depths, the x-axis represents sequencing depth, and the y-axis shows the proportion of bases achieving that specific depth. The distribution typically follows a Poisson pattern around the average depth. The right figure displays cumulative base proportions at various depths, where the x-axis represents sequencing depth, and the y-axis shows the cumulative proportion of bases with at least that depth. For example, a sequencing depth of 0 corresponds to a 100% base proportion, indicating that 100% of bases have a sequencing depth greater than 0.

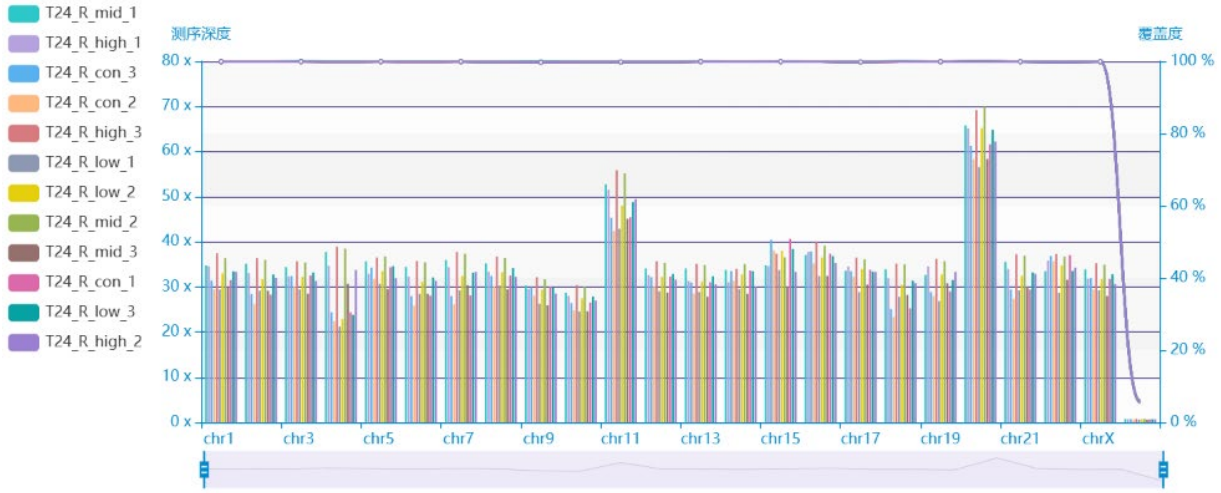

Figure 9 Coverage depth (left) and coverage (right) of each chromosome

The horizontal axis indicates chromosome numbering, while the left vertical axis shows average coverage depth and the right vertical axis displays coverage rate. For calculating coverage depth per chromosome, the formula is: sequencing data volume per chromosome / total length of exon regions on each chromosome. When determining coverage rate, the formula is: total length covered by each chromosome / total length of exon regions on each chromosome.

3.6 Coverage statistics

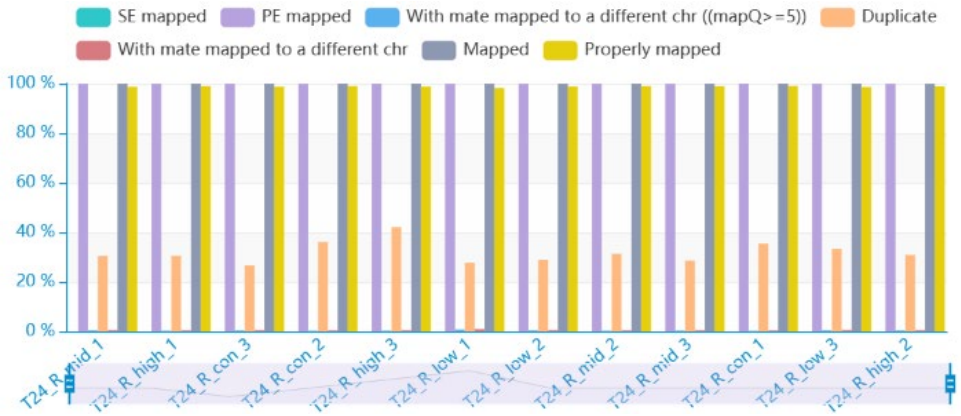

Figure 10 Distribution of sequencing reads across different genomic features

3.7 SNP statistical results

A human genome typically contains approximately 3.6 million SNPs. Over 95% of these high-frequency SNPs (with allele frequencies exceeding 5% in the population) are documented in the dbSNP database (Sherry et al., 2001). The transversion-to-crossover ratio (Ts:Tv) serves as an indicator of SNP dataset accuracy, averaging around 2.2 across the genome and approximately 3.2 in coding regions. The detection and statistical analysis results for SNPs are presented as follows:

Experiment time: 2025-06-29, 18:00 - 2025-06-30, 21:30

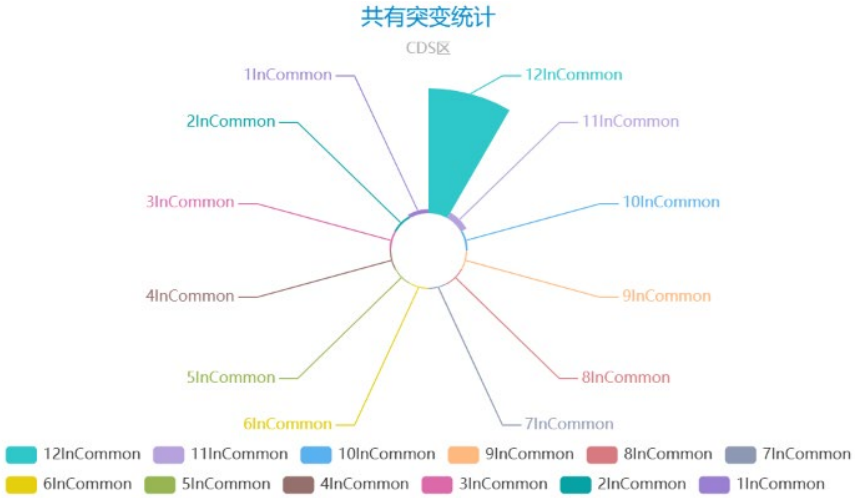

Figure 11 Statistics of SNP mutations in the CDS region (3InCommon indicates the number of mutations shared among 3 samples)

Based on the alignment results, we identified SNP sites using bcftools (Li et al., 2009) and filtered them according to internationally recognized standards. In this sequencing sample, an average of 3,155,367 SNP sites were detected per sample, with 198,020 in exonic regions, 1,718,276 in intergenic regions, and 546 in splice regions. Among coding region SNPs, each sample contained 10,017 synonymous mutations, 9,245 missense mutations, 69 stop-gain SNPs, and 10 stop-loss SNPs. The detailed statistical results are presented in the table below:

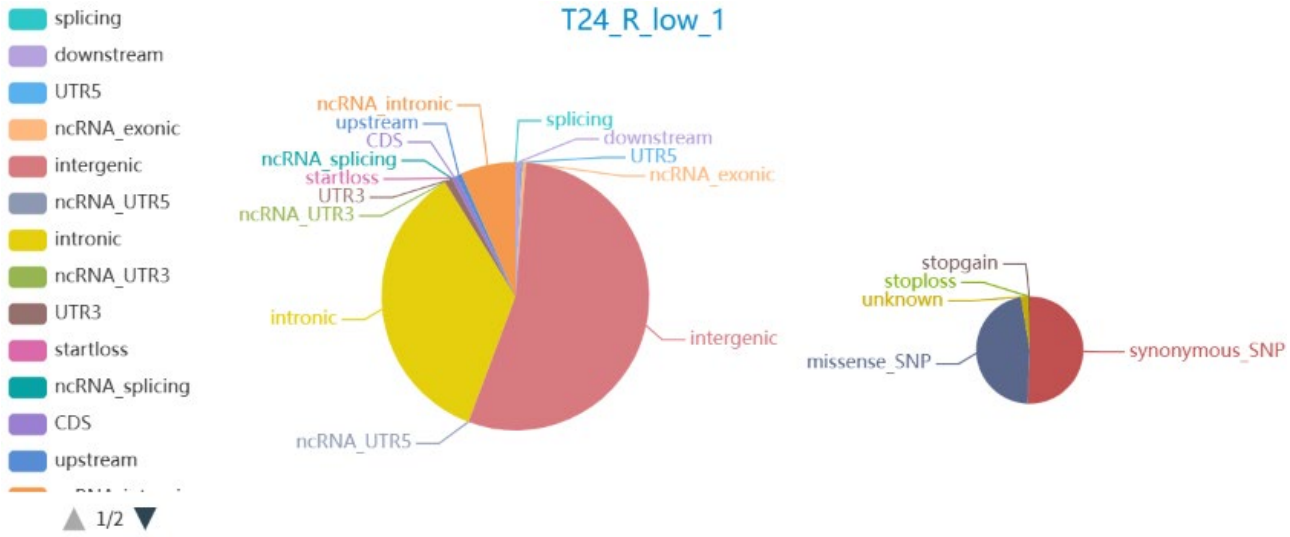

Figure 12 Distribution of SNP numbers in different regions of genome (left) and different types of SNP numbers in coding region (right)

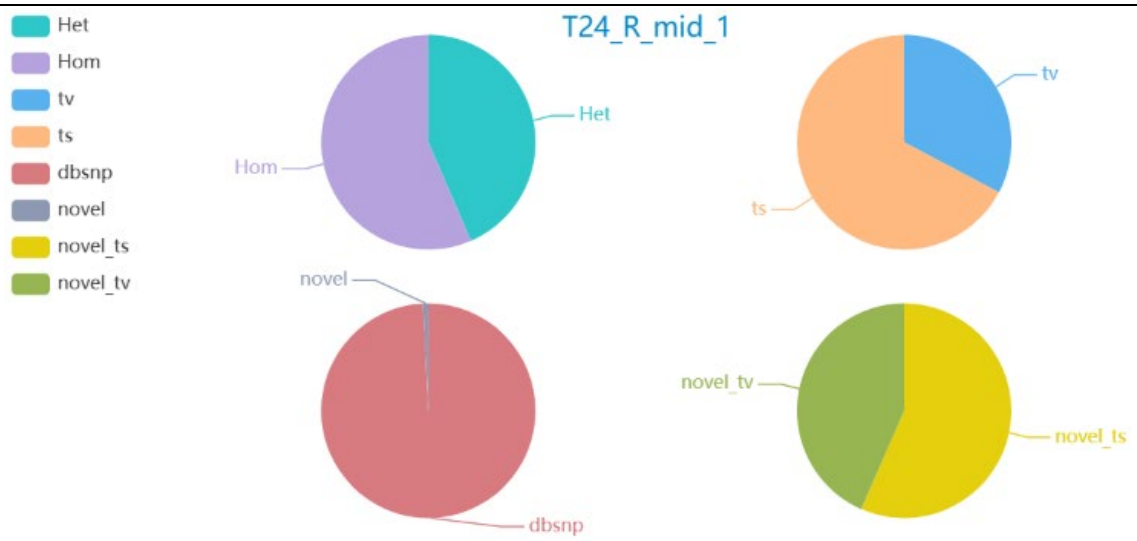

Figure 13 Genomic SNP characteristics

3.8 InDel statistical results

A human genome typically contains approximately 350,000 InDels (Insertions and Deletions), which are DNA insertions or deletions shorter than 50bp. These insertions or deletions in coding regions or splice sites may alter protein translation. Transcoding variants, characterized by base sequences of insertion or deletion lengths that are non-integer multiples of three, can lead to complete frame changes. Compared to non-transcoding variants, transcoding variants face greater regulatory constraints.

Based on the alignment results, we identified InDels using bcftools (Li et al., 2009) and filtered the results according to internationally accepted standards. The sequencing samples contained an average of 923,236 InDels per sample, with 571 in exonic regions, 468,933 in intergenic regions, and 385 in splice sites. Among coding region InDels, each sample averaged 68 frameshift deletions, 61 frameshift insertions, 177 non-frameshift deletions, 149 non-frameshift insertions, 3 InDels causing loss of stop codons, and 10 InDels leading to gain of stop codons. The statistical results of InDels are as follows:

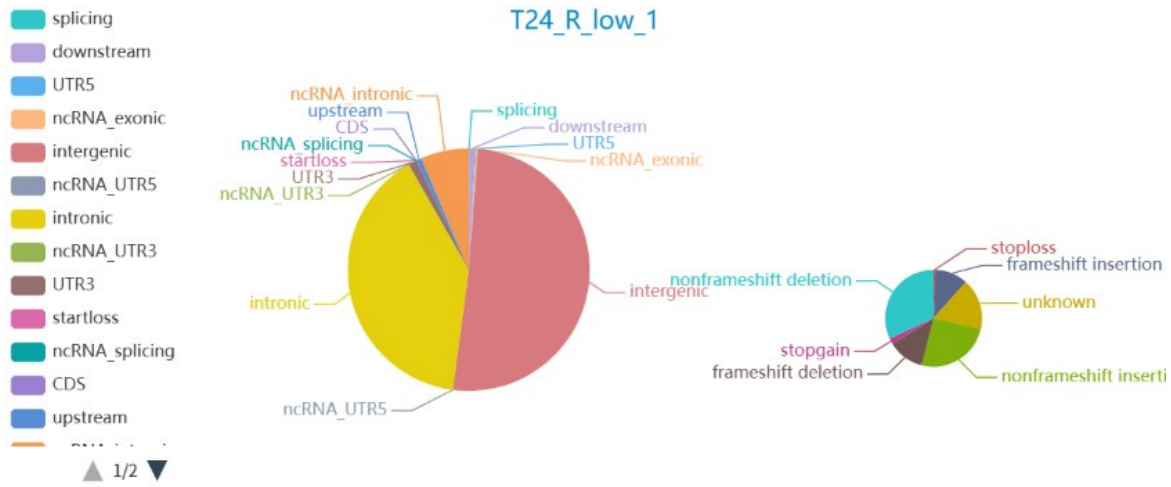

Figure 14 Distribution of InDel numbers in different regions of the genome (left) and different types of InDel numbers in coding regions (right)

Figure 15 Genomic InDel features

3.9 Somatic cell variation detection (PON)

Somatic mutations refer to genetic alterations acquired in tumor cells (or drug-resistant derivatives in this context) that are not present in the germline of the organism. Unlike parental inheritance, these mutations neither originate from parents nor are passed to offspring, yet they can alter the genetic makeup of specific cells in the current generation. Particularly, driver mutations within somatic mutations play a pivotal role in explaining tumor development and progression. Additionally, the emergence of drug resistance in tumors is also associated with somatic mutations. Therefore, focusing on somatic mutations constitutes the core focus of tumor genomics research, distinguishing it from conventional disease studies through this unique characteristic.

SNVs (Single Nucleotide Variants) are genetic variations caused by single nucleotide substitutions in the genome. We primarily use the PON mode (https://gatk.broadinstitute.org/hc/en-

us/articles/360036364532-CreateSomaticPanelOfNormals-BETA-) in the muTect2 software to identify somatic SNV sites.

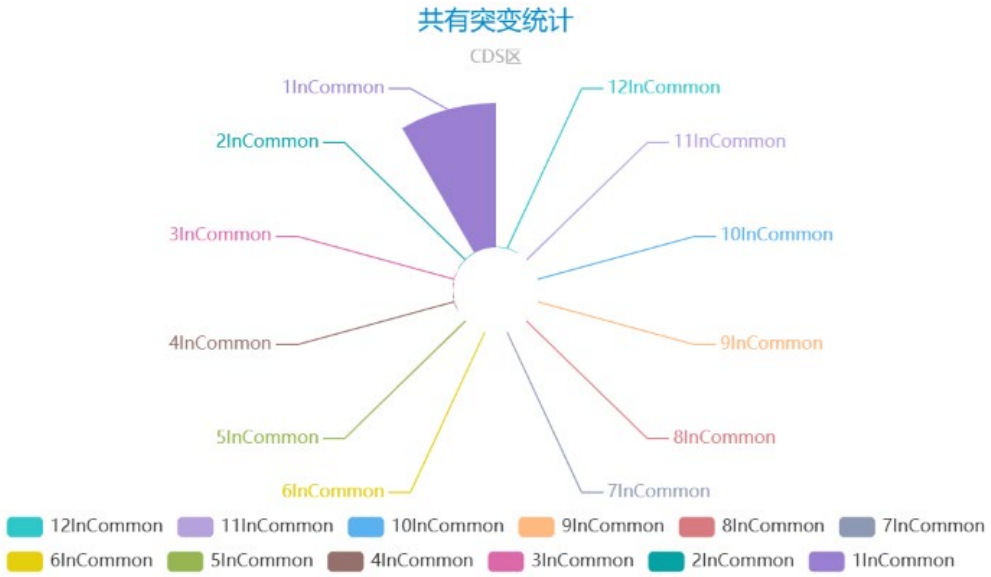

Figure 16 Somatic SNV mutation statistics in the CDS region (3InCommon indicates mutations present in all three samples)

IV. Results analysis

High-frequency mutated genes and genomic structural variations identified in drug-resistant cell lines suggest they may play a regulatory role in the development of RC48 resistance. By comparing mutation profiles between parent strains and cell lines with varying resistance levels, we observe a progressive enrichment trend of drug-related genes, demonstrating a significant correlation between genomic dynamic evolution and drug-resistant phenotypes.

Functional mutations in key genes (such as kinase domain or protein domain variations) may directly or indirectly enhance drug resistance by affecting the binding efficiency of drug targets or downstream signaling capabilities. Additionally, copy number variations (e.g., amplification or deletion of specific genes) can alter the activity of critical signaling pathways, promoting cell survival or apoptosis evasion, thereby forming drug-resistant phenotypes.

Structural variations and large-scale rearrangements detected in drug-resistant cell lines may accelerate the accumulation of resistance-related mutations by driving genomic instability. These variations could disrupt the functions of genes involved in cell cycle regulation, DNA repair, or epigenetic modification, thereby providing a genetic basis for the adaptive evolution of drug-resistant clones.

Experiment time: 2025-06-29, 18:00 - 2025-06-30, 21:30

Whole-genome variation data suggest that drug resistance development likely involves coordinated interactions across multiple genes and signaling pathways. For example, combined modifications in target genes and activation of survival signaling pathways may collectively diminish therapeutic efficacy, while dysregulation of DNA repair-related genes could further enhance the screening pressure tolerance of drug-resistant clones.

The biological significance of genomic data requires integrated validation through the combination of transcriptomic, proteomic, and functional experiments to clarify the specific roles of candidate genes in drug resistance. Furthermore, the evolutionary patterns of drug resistance-related variants and their clinical relevance still need further exploration in more complex models or clinical samples.

In conclusion, whole-genome sequencing has systematically revealed genomic variation patterns in the drug-resistant cell lines, providing crucial candidate genes and molecular clues for deciphering the mechanisms of RC48 acquired resistance. Future research should focus on validating the biological functions of key variants and conducting integrated multi-omics analyses to comprehensively elucidate the network-regulated regulatory patterns underlying antibiotic resistance mechanisms.