Electronic Supplementary Material (ESI) for Molecular BioSystems.

Supplementary Information

RNA-sequencing dissects the transcriptome of polyploid cancer cells that resistant to combined treatments of cisplatin with paclitaxel and docetaxel

Qianqian Wang, Fei Lu, and Rongfeng Lan

1. Procedures of RNA sequencing

a. RNA sample collection and preparation

Total cellular RNA was extracted using RNAiso Plus. RNA degradation and contamination was monitored on 1% agarose gels. And the purity of RNA was examined using the NanoPhotometer® spectrophotometer (IMPLEN, CA, USA). RNA concentration was measured using Qubit® RNA Assay Kit in Qubit® 2.0 Flurometer (Life Technologies, CA, USA). RNA integrity was assessed using the RNA Nano 6000 Assay Kit of the Bioanalyzer 2100 system (Agilent Technologies, CA, USA).

b. Sequencing library preparation

A total amount of 3 μg RNA per sample was used as input for the RNA sample preparations. Sequencing libraries were generated using NEBNext® Ultra™ RNA Library Prep Kit for Illumina® (NEB, USA) following manufacturer's protocols and index codes were added to attribute sequences to each sample. Briefly, mRNA was purified from total RNA using poly-T oligo-attached magnetic beads. Fragmentation was carried out using divalent cations under elevated temperature in NEBNext First Strand Synthesis Reaction Buffer (5X). First strand cDNA was synthesized using random hexamer primer and M-MuLV Reverse Transcriptase (RNase H-). Second strand cDNA synthesis was subsequently performed using DNA Polymerase I and RNase H. Remaining overhangs were converted into blunt ends via exonuclease/polymerase activities. After adenylation of 3' ends of DNA fragments, NEBNext Adaptor with hairpin loop structure were ligated to prepare for hybridization. In order to select cDNA fragments of preferentially 150~200 bp in length, the library fragments were purified with AMPure XP system (Beckman Coulter, Beverly, USA). Then 3 μl USER Enzyme (NEB, USA) was used with size-selected, adaptor-ligated cDNA at 37 °C for 15 min followed by 5 min at 95 °C before PCR. Then PCR was performed with Phusion High-Fidelity DNA polymerase, Universal PCR primers and Index (X) Primer. At last, PCR products were purified (AMPure XP system) and library quality was assessed on the Agilent Bioanalyzer 2100 system.

c. Clustering and sequencing

The clustering of the index-coded samples was performed on a cBot Cluster Generation System using TruSeq PE Cluster Kit v3-cBot-HS (Illumia) according to the manufacturer's instructions. After cluster generation, the library preparations were sequenced on an Illumina Hiseq platform and 125 bp/150 bp paired-end reads were generated.

2. Data Analysis

a. Quality control

Raw data (raw reads) of fastq format were firstly processed through in-house perl scripts. Clean data (clean reads) were obtained by removing reads containing adapter, reads containing ploy-N and low quality reads from raw data. At the same time, Q20, Q30 and GC content the clean data

were calculated. All the downstream analyses were based on the clean data with high quality.

b. Reads mapping

Reference genome and gene model annotation files were downloaded from genome website directly. Index of the reference genome was built using Bowtie v2.2.3 and paired-end clean reads were aligned to the reference genome using TopHat v2.0.12(Ref S1). TopHat was selected as the mapping tool for that TopHat can generate a database of splice junctions based on the gene model annotation file and thus a better mapping result than other non-splice mapping tools.

c. Quantification of gene expression level

HTSeq v0.6.1 was used to count the reads numbers mapped to each gene. And then FPKM of each gene was calculated based on the length of the gene and reads count mapped to this gene. FPKM, expected number of Fragments Per Kilobase of transcript sequence per Millions base pairs sequenced, considers the effect of sequencing depth and gene length for the reads count at the same time, and is currently the most commonly used method for estimating gene expression levels (Ref S2, S3).

d. Differential expression analysis

Differential expression analysis of two biological replicates per condition) was performed using the DESeq R package (1.18.0) (Ref S4). DESeq provide statistical routines for determining differential expression in digital gene expression data using a model based on the negative binomial distribution. The resulting p-values were adjusted using the Benjamini and Hochberg's approach for controlling the false discovery rate. Genes with an adjusted p-value <0.05 found by DESeq were assigned as differentially expressed.

e. Clustering analysis of DEGs

Clustering analysis of DEGs is used to determine the expression pattern of genes under different experimental conditions. The genes with same or similar expression pattern will sorted into a cluster, in order to identify the unknown function of genes or functions of unknown genes. Taking the FPKM value for the gene expression level, hierarchical clustering analysis is performed. Different color represents of different clusters, while the same cluster indicates similar gene expression pattern, which might have similar functions or participate in the same biological process.

K-means cluster was realized based on the relative expression level of DEGs ($\log_2$(Fold Change)) operating by clustering software package pheatmap. Clustering is achieved by distance algorithm that is calculating the distance of each gene, and then followed by iterative calculation of the relative distance between genes, finally obtained different subclusters according to the relative distance of the gene. Software is available on http://bonsai.hgc.jp/~mdehoon/ software/cluster/s

oftware.htm

f. GO and KEGG enrichment analysis of DEGs

Gene Ontology (GO) enrichment analysis of differentially expressed genes was implemented by the GOseq R package (Ref S5), in which gene length bias was corrected. GO terms with corrected p-value less than 0.05 were considered significantly enriched by differential expressed genes. KOBAS software was used to test the statistical enrichment of differential expression genes in KEGG pathways (Ref S6, 7).

g. References

S1 Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq.

Bioinformatics 2009;25(9):1105-1111.

S2 Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* 2015;31(2):166-169.

S3 Trapnell C, Williams BA, Pertea G*, et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 2010;28(5):511-515.

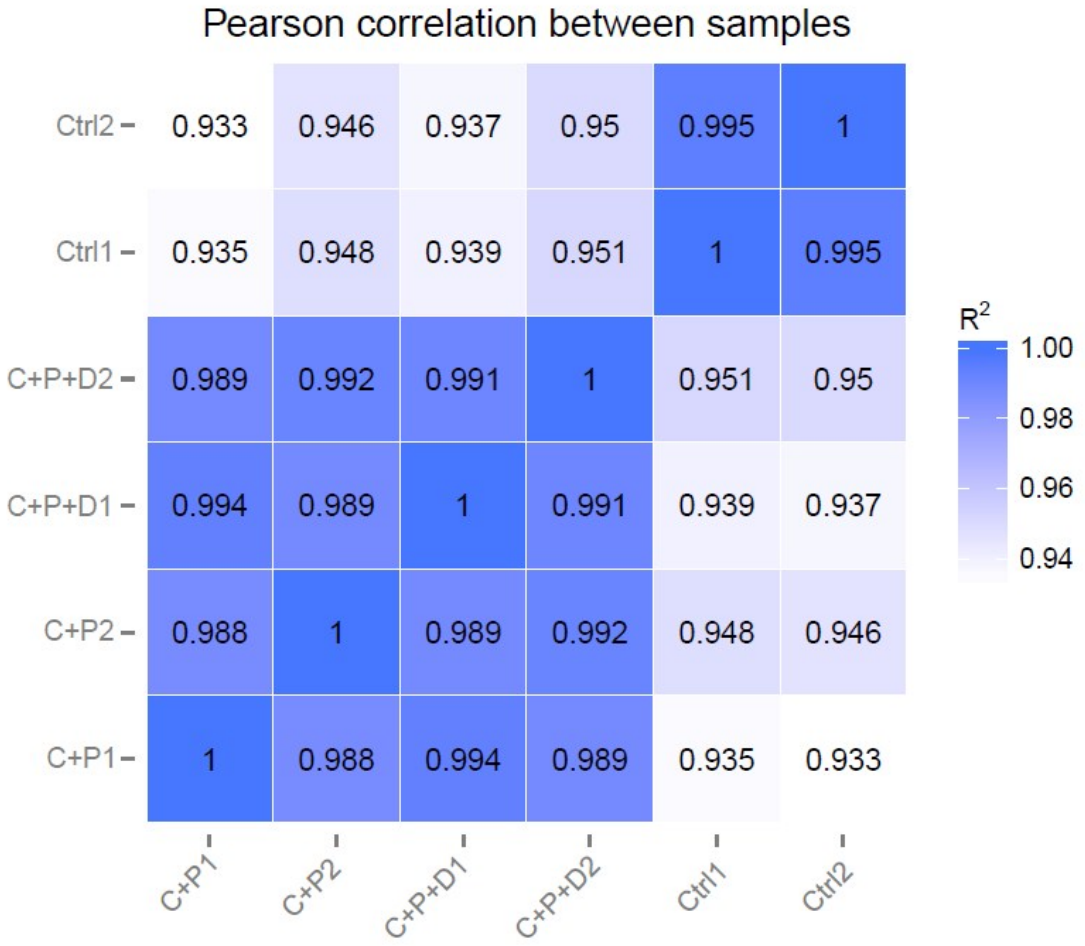
S4 Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol* 2010;11(10):R106.

S5 Young MD, Wakefield MJ, Smyth GK*, et al.* Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol* 2010;11(2):R14.

S6 Kanehisa M, Araki M, Goto S*, et al.* KEGG for linking genomes to life and the environment. *Nucleic Acids Res* 2008;36(Database issue):D480-484.

S7 Mao X, Cai T, Olyarchuk JG*, et al.* Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. *Bioinformatics* 2005;21(19):3787-3793.

3. Pearson correlation analysis of samples for RNA-Sequencing.



The correlation between gene expression levels among samples is an important indicator to test the reliability and sample selection. The closer the correlation coefficient is to 1, the higher the similarity of the gene expression patterns between samples. Normally the square of the Pearson

correlation coefficient (R^2) should be greater than 0.92 (ideal for sampling and experimental conditions).

4. Primers used in RT-PCR.

p53 signaling pathway (up regulated)

Gene name	Primers
Gadd45a	F: 5′- CAGAAGACCGAAAGGATGGA-3′ R: 5′- GCACGGATGAGGGTGAAATG-3′
Cdkn1a	F: 5′- ACGCCGCTGGAGGGCAACTT-3′ R: 5′- CTTCAGGCCGCTCAGACACC-3′
Ccnd1	F: 5′- GGAGCAGAAGTGCGAAGA-3′ R: 5 - GGGCCGGATAGAGTTGTC-3′
Mdm2	F: 5′- GGGGAAAGATAAAGTGGA-3′ R: 5′- ATGGTTGGGAATAGTCGT-3′
Pten	F: 5′- GCTGGAAAGGGACGGACTG-3′ R: 5′- GCCACGGGTCTGTAATCC-3′
Rchy1	F: 5′- TGCTTGGAGGACATTCAC-3′ R: 5′- TTCGGATGGCATGGGAGT-3′
Ccng1	F: 5′- GTCTGCGGCTTGAAACTA-3′ R: 5′- AAGATGCTTCGCCTGTAC-3′
Siah1a	F: 5′- TGCCTCTTCTGGATGTGA-3′ R: 5′- GCAGGGTGGTAATGGACT-3′
Ppm1d	F: 5′- TGCTTCGGGCAGATAACA-3′ R: 5′- CGGTGACTTGATTGGTGGT-3′
Trp73	F: 5′- ATCTCCATCGGCGGCTCT-3′ R: ‘5- GGCTGCTTACGGGACTTG-3′

Chronic myeloid leukemia (up regulated)

Gene name	Primers
Shc1	F: 5′- TCAGGAATCCACCGAAGC-3′ R: 5′- AAGTGGCTGGACATCTGG-3′
Shc3	F: 5′- TCCACCGCCAGCCTGAAT-3′ R: 5′- GCCCGTCACAGCATTCCA-3′
Araf	F: 5′- CCTCCACGCCAAGAACAT-3′

	R: 5’- CCTGCATACGGATCACCT-3’
Map2k1	F: 5’- TGGGCACGAGATCCTACA-3’
	R: 5’- CTGCGTCTCCTTCCACAT-3’
Rela	F: 5’- CCCCTTTCACGTTCCTAT-3’
	R: 5’- CCCAGAGTTCCGGTTTAC-3’
Ikbkg	F: 5’- GATTCGAGCAGTTAGTGAGC-3’
	R: 5’- TGGAAGAGCTGGTGATAGG-3’
Runx1	F: 5’- TGGCAGGCAACGATGAAA-3’
	R: 5’- GCTCTATGGTAGGTGGCAACT-3’
Smad3	F: 5’- CTGGCTACCTGAGTGAAGATG-3’
	R: 5’- TGTGAGGCGTGGAATGTC-3’

DNA replication (down regulated)

Gene name	Primers
Pcna	F: 5’- AAGAAGAGGAGGCGGTAA-3’
	R: 5’- AGTGTCCCATGTCAGCAA-3’
Mcm2	F: 5’- CACAGAGCCCATCATTTCC-3’
	R: 5’- GCCACCATTAGTCAACCCT-3’
Rfc1	F: 5’- AGACGCACTTGTACGACC-3’
	R: 5’- CCGCTTTCACCTTGGGAT-3’
Dna2	F: 5’- ACGATGCGAAGGATACGG-3’
	R: 5’- AGCAGAACGCTGAACCCA-3’
Fen1	F: 5’- ACCAAGAGGCTCGTGAAG-3’
	R: 5’- GGCAGTTAAGTGTCGCAT-3’
Prim1	F: 5’- AGCCCTTGTTCCTGAGAC-3’
	R: 5’- CTCCCTTGCTGACGTTGA-3’
Rpa3	F: 5’- GCCACTTGACGAGGAAAT-3’
	R: 5’- CATGTTGTGGAAGCCCTA-3’

Cell Cycle (down regulated)

Gene name	Primers
Orc1	F: 5’- GCGGATTAGGCAACAGCTT-3’
	R: 5’- GGCACGATGAGGTGTTCTA-3’

Skp2	F: 5’- GAGGTGGACAGTGAGAACA-3’
	R: 5’- GAGGCACAGACAGGAAAA-3’
Cdc7	F: 5’- AGACCACAGCGATTGACA-3’
	R: 5’- CTGGGACTTCTTTGCTACAC-3’
Dbf4	F: 5’- AGAGCCCACAACCTATTC-3’
	R: 5’- TCCACATGACTGCGACA-3’
Cdkn2c	F: 5’- GCTTCTCCTCAGAGGTGCTA-3’
	R: 5’- AGGGAGGTGGCCTTCTTT-3’
Bub1	F: 5’- GCATCTTTACCCTGTCCT-3’
	R: 5’- CAAGTGTCTGCACGCATT-3’
Mad2l1	F: 5’- TCCCAGAAAGCCATACAG-3’
	R: 5’- GTAGACGGACTTCTTCACA-3’
Cdc20	F: 5’- GGAGGTGACCGCTTTATC-3’
	R: 5’- TGAGCCTGAGGATCTTGG-3’
Ttk	F: 5’- TACCAGAAAGCCCGAGTG-3’
	R: 5’- GACAGGCAGGTGGAAAGT-3’
Plk1	F: 5’- GCCTAAGTCTTTGCTGCTC-3’
	R: 5’- AGTGCCTTCCTCCTCTTGT-3’
Cdc6	F: 5’- TGGCTAAGCAACTCCCGAT-3’
	R: 5’- TCGCAGCACTGTCCAGAA-3’

5. siRNA target sequences and RT-PCR primers for representative DEGs.

Gene name	siRNA and RT-PCR Primers
Abcb9 (Homo sapiens)	siRNA: CCTCTTCGTGGGCATCTAT
	RT-PCR primers F: 5’- CGGGTGGACTTTGAGAAT -3’
	R: 5’- TGGTCGTAGGCGCTGAT -3’
Foxm1 (Homo sapiens)	siRNA: GCCAACCGCTACTTGACAT
	RT-PCR primers F: 5’- ACCACGGGTCAGCTCATA -3’
	R: 5’- CGCCACTAAAGAACTTACTC -3’
Slc10a3 (Homo sapiens)	siRNA: GCACAGGTCCCTTAAGCAT
	RT-PCR primers F: 5’- CTGCTGTTCATTGCCATCC -3’
	R: 5’- CGTGATACCCACCAGTACGAT -3’
Slc2a9 (Homo sapiens)	siRNA: GCTGGCCAATAATGGGTTT

RT-PCR primers	F: 5’- TGACTGCCATCTTTATCTGC -3’	
	R: 5’- AGCTCTTGCCTCGTTGTG -3’	