9057

Genomics and Proteomics 2

Data analysis report

Word count: 1987

Electron acceptors promote apoptosis, cancer progression and various other cellular processes

Abstract

Electron acceptors can lead to oxidative stress and distinct cellular changes. It has been a contradiction of whether the electron acceptor promotes or inhibits cancer progression (Gorrini *et al.*, 2013). Here, RNA-seq was performed on K562 cancer cell lines treated with electron acceptors including alpha-ketobutyrate (AKB) or pyruvate. Comparing the transcriptome of AKB or pyruvate treated K562 cells with the untreated group, I aimed to explore the global transcriptional response to the electron acceptors in K562 cells. The results show that there are enrichments on cancer progression characteristics including increased oncogene expression and metastasis capability. Electron acceptors also induce many other cellular changes including apoptosis, RNA processing and metabolic process as well as organic acid metabolic process. Moreover, two different electron acceptors may lead to opposite cellular changes.

Introduction to the scientific question

K562 cell line is a type of chronic myelogenous leukemia (CML) cell line taken from the bone marrow (Koeffler and Golde, 1980). After treating K562 cells with electron acceptors, RNA was extracted and sequenced using next-generation sequencing. The data set contain 3 groups (control, AKB-treated and pyruvate-treated) with 4 replicates in each group. The experiment aims to find out the global transcriptional response to electron acceptors in K562 cells. Since electron acceptors can induce oxidative stress and DNA damage, it is suspected that they can induce apoptosis and promote cancer progression. Electron acceptors may also induce other cellular changes since they have broad effects on many cellular processes and components. Moreover, whether two different electron acceptors exert similar effects remains to be investigated.

Analysis methods

The reference genome (fasta) and gene annotation (gtf) files were downloaded from the Ensembl reference genome database (Table S4). I used GRCh38 for analysis since GRCh38 is an improved version of the human genome where many gaps were closed, sequencing errors corrected, and centromere sequences modelled compared to GRCh37.

Firstly, I performed the quality control for all raw sequencing files using FastQC. Then, Cutadapt was used to trim out the low-quality bases. Cleaned fastq files were aligned to the reference genome using HISAT2 with default settings. HISAT2 is a splice-aware alignment tool and is widely used with more than 4000 citations. Also, it requires smaller memory compared to STAR. After alignment, the count matrix was generated from the aligned read files (sam) using featureCounts. featureCounts is a widely used and reliable software that can assign sequence reads to genomic features (Liao *et al.*, 2014). The normalization of the raw count matrix and calling of differentially expressed genes (DEGs) were performed using R package DESeq2. Among differential expression tools, the results generated by DESeq2 and statistical tests are clustered together which indicates that DESeq2 produces reliable results (Schurch *et al.*, 2016). DESeq2 is also actively maintained by the developers. AKB and pyruvate groups were compared to the control group respectively, which generates two sets of DEGs. Next, gene ontology and gene set enrichment analysis were performed to analyze the DEGs. These analyses were performed using clusterProfiler package. clusterProfiler supports all the gene signatures downloaded from MSigDB database. MSigDB is a reliable source of gene signatures since it is constantly maintained and updated by the government and top universities/institutes.

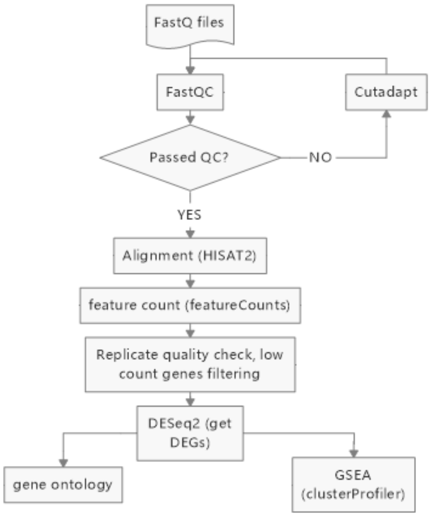


Figure 1. The workflow of data processing

Results

Overall quality of the raw sequencing data The quality of all data is similar. All data passed the “per base sequence quality”, although there is a little drop at the end of the reads which is due to the next generation sequencing deficiencies. The data does not have an overrepresented sequence and have no adaptor content. All data failed to pass the quality control of “per base sequence content” and “sequence duplication levels” (Figure 2a, 2b). I trimmed the first 14 bases to solve the first problem (Figure 2c). The second problem may be caused by PCR duplication or some highly expressed genes. I decided not to treat this problem since it may underestimate the expression level of some genes. Some data also have problem with “Per sequence GC content” (Figure 2d). However, the problem is not serious, and I will proceed with the analysis.

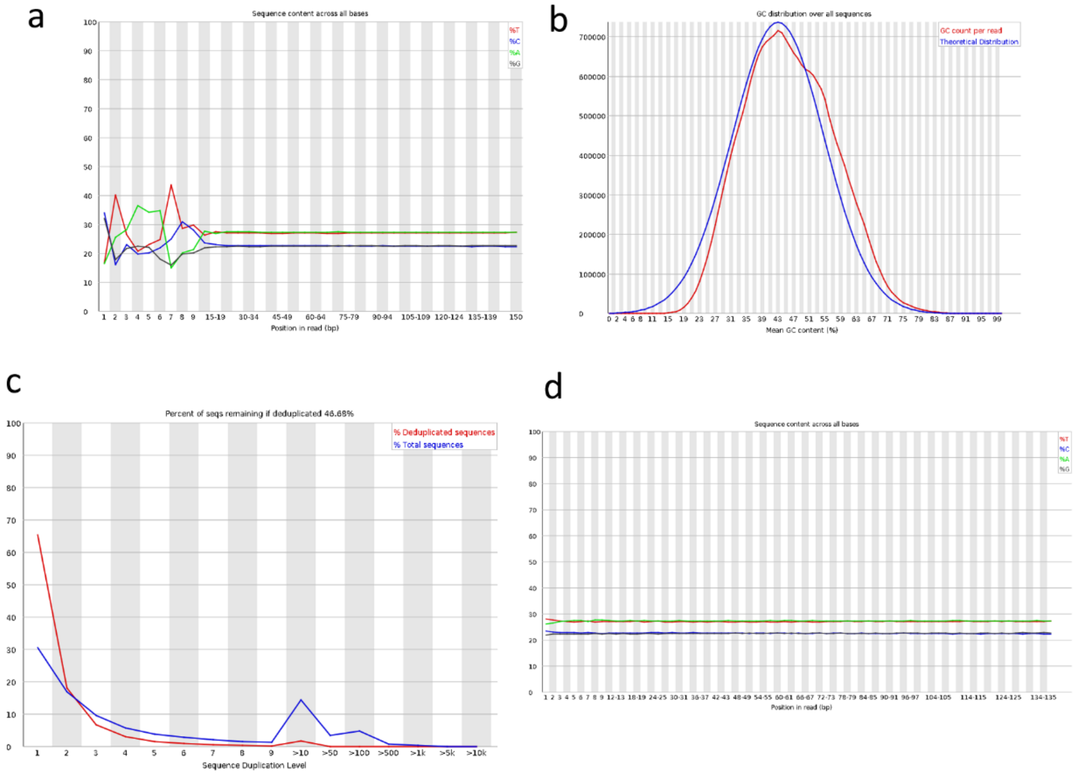
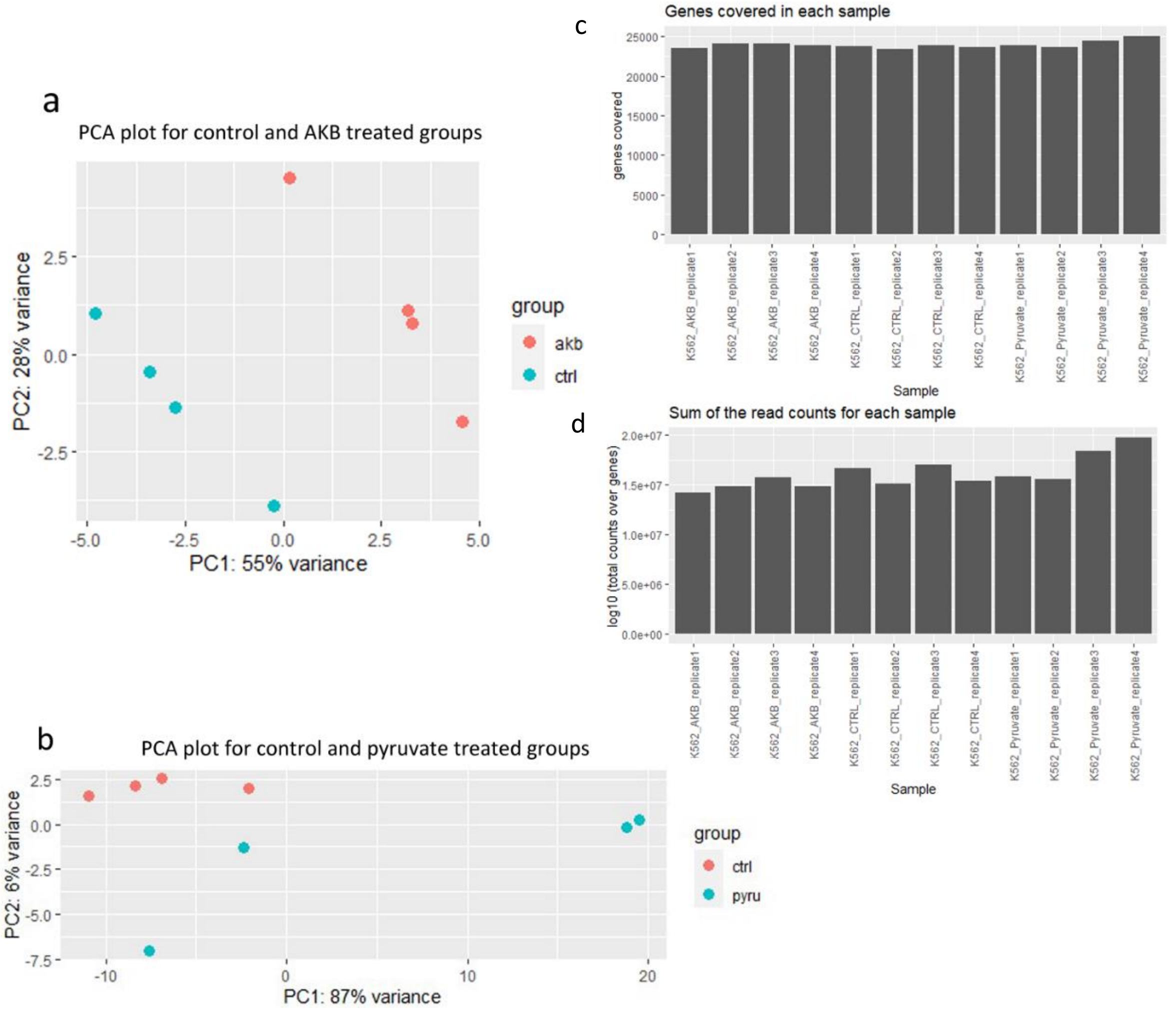


Figure 2. Quality control for raw sequencing data (FastQC results of ctrl-1_FRAS202101646-1r_1.clean.fq.gz shown) **a** “Per base sequence content” before trimming. **b** “Per sequence GC content” before trimming. **c** “Sequence duplication level”. **d** “Per base sequence content” after trimming.

Mapping statistics The overall alignment rates of all samples are around reasonable 97% (Table S2). All samples have less than 5% multi-aligned reads which will not affect the following counting step (Pertea et al., 2016). Around 70% aligned reads were assigned successfully to the genes in the feature count step which is also reasonable (Table S3).

Quality of the replicate samples To assess the quality of the replicate samples, we first plot the total genes covered and the sum of the read counts in every sample. In all 12 samples, more than 20000 genes have at least one count and the numbers of genes in all samples are similar (Figure 3c). The sums of read counts are similar among all samples (Figure 3d).

Clustering (PCA) and sample distance heatmap were plotted to assess the overall similarity between samples. Before PCA and sample distance, vst normalization was applied to the read counts of all samples to make the data homoscedastic. PCA results for both AKB and pyruvate groups show that the same replicates from different groups are clustered to each other well (Figure 3a, 3b), although there is variation among pyruvate replicates. Sample distance heatmap shows that except for control replicate 4, all the other groups are clustered quite well (Figure 3e, 3f). There is no evidence of a potential problem with the wet lab experiment. Thus, there are no clear and justifiable reasons for excluding control replicate 4. I will proceed with the analysis with all replicates.



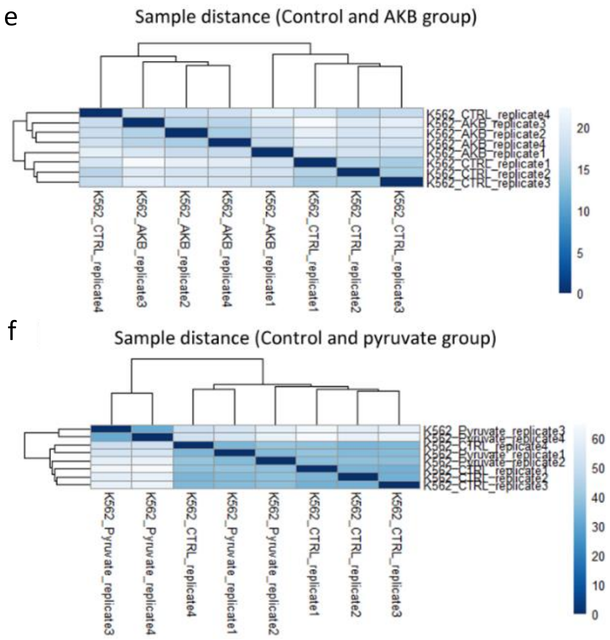
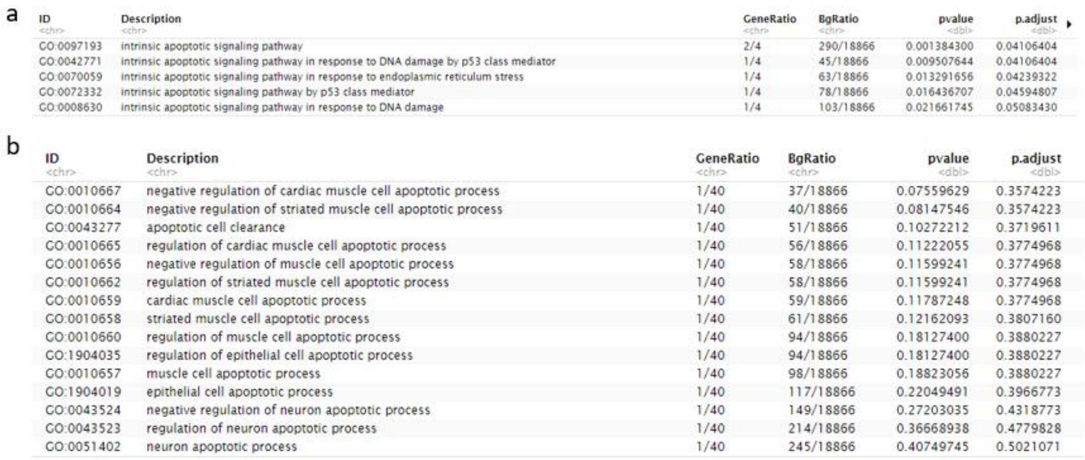


Figure 3. The quality of replicate samples. **a-b** PCA plots after vst normalization. **c** total number of genes covered in each sample. **d** sum of the read counts in each sample. **e-f** Sample distance heatmap after vst normalization.

Differentially expressed genes Firstly, genes with less than 15 total read counts (across all 12 samples) are excluded. Then, the differentially expressed genes were called using DESeq function in DESeq2 package. In AKB group, 1710 genes (837 log2 fold change > 0) were identified as DEGs with a 5% false discovery rate. In pyruvate group, 653 genes (443 log2 fold change > 0) were identified as DEGs with a 5% false discovery rate. This might suggest that AKB brings more potent effects on K562 cells. Among these genes, only a few genes (less than 10%) have log2 fold change > 2 or < -2. Next, enrichment and network analysis were performed to analyze the DEGs.

Gene ontology (GO) analysis reveals different biological processes induced by AKB and pyruvate

All DEGs with log2 fold change > 1 were selected to perform gene ontology analysis. Their gene names were imported into function enriGO from clusterProfiler. The result shows that in AKB groups, genes are enriched in apoptotic signalling pathways (Figure 4a). Most of the enrichments have adjusted p-values lower than 0.05. This supports the hypothesis that electron acceptors induce apoptosis. However, in the pyruvate group, there is no significance in all apoptotic processes.


a

ID	Description	GeneRatio	BgRatio	pvalue	p.adjust
<chr>	<chr>	<chr>	<chr>	<dbl>	<dbl>
GO:0097193	intrinsic apoptotic signaling pathway	2/4	290/18866	0.001384300	0.04106404
GO:0042771	intrinsic apoptotic signaling pathway in response to DNA damage by p53 class mediator	1/4	45/18866	0.009507644	0.04106404
GO:0070059	intrinsic apoptotic signaling pathway in response to endoplasmic reticulum stress	1/4	63/18866	0.013291656	0.04239322
GO:0072332	intrinsic apoptotic signaling pathway by p53 class mediator	1/4	78/18866	0.016436707	0.04594807
GO:0008630	intrinsic apoptotic signaling pathway in response to DNA damage	1/4	103/18866	0.021661745	0.05083430

b

ID	Description	GeneRatio	BgRatio	pvalue	p.adjust
<chr>	<chr>	<chr>	<chr>	<dbl>	<dbl>
GO:0010667	negative regulation of cardiac muscle cell apoptotic process	1/40	37/18866	0.07559629	0.3574223
GO:0010664	negative regulation of striated muscle cell apoptotic process	1/40	40/18866	0.08147546	0.3574223
GO:0043277	apoptotic cell clearance	1/40	51/18866	0.10272212	0.3719611
GO:0010665	regulation of cardiac muscle cell apoptotic process	1/40	56/18866	0.11222055	0.3774968
GO:0010656	negative regulation of muscle cell apoptotic process	1/40	58/18866	0.11599241	0.3774968
GO:0010662	regulation of striated muscle cell apoptotic process	1/40	58/18866	0.11599241	0.3774968
GO:0010659	cardiac muscle cell apoptotic process	1/40	59/18866	0.11787248	0.3774968
GO:0010658	striated muscle cell apoptotic process	1/40	61/18866	0.12162093	0.3807160
GO:0010660	regulation of muscle cell apoptotic process	1/40	94/18866	0.18127400	0.3880227
GO:1904035	regulation of epithelial cell apoptotic process	1/40	94/18866	0.18127400	0.3880227
GO:0010657	muscle cell apoptotic process	1/40	98/18866	0.18823056	0.3880227
GO:1904019	epithelial cell apoptotic process	1/40	117/18866	0.22049491	0.3966773
GO:0043524	negative regulation of neuron apoptotic process	1/40	149/18866	0.27203035	0.4318773
GO:0043523	regulation of neuron apoptotic process	1/40	214/18866	0.36668938	0.4779828
GO:0051402	neuron apoptotic process	1/40	245/18866	0.40749745	0.5021071

Figure 4. Gene ontology analysis of apoptotic process. Apoptotic processes enriched in DEGs with log2 fold change > 1, adjusted p-value < 0.05 **a** in AKB group and **b** in pyruvate group.

Another function named “gseGO” in the clusterProfiler package can perform gene set enrichment analysis using gene ontology (Yu et al., 2012). It uses all DEG information instead of traditional gene ontology analysis which only uses genes with high fold change. By importing ranked gene list (based on log2 fold change) to gseGO function, I identified that the DEGs are enriched in some ontologies. In the AKB group, DEGs are enriched in RNA processing and metabolic process (Figure 5a). In the pyruvate group, DEGs are enriched organic acid metabolic process and transmembrane transport (Figure 5b).

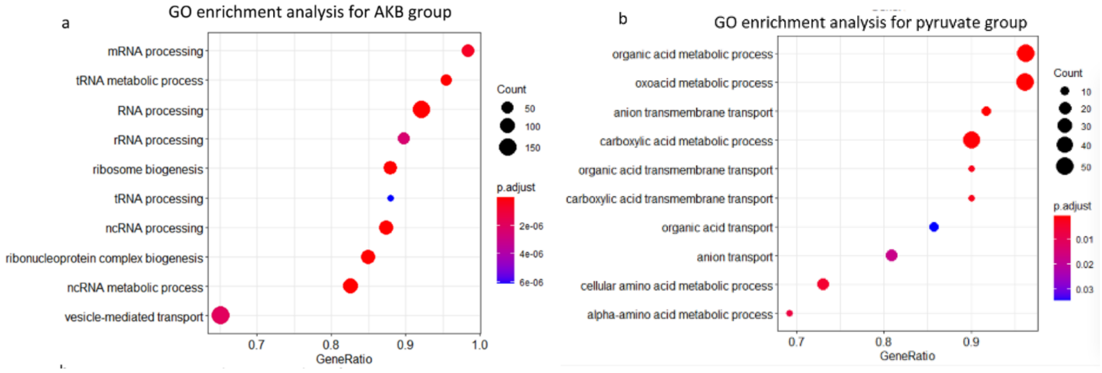
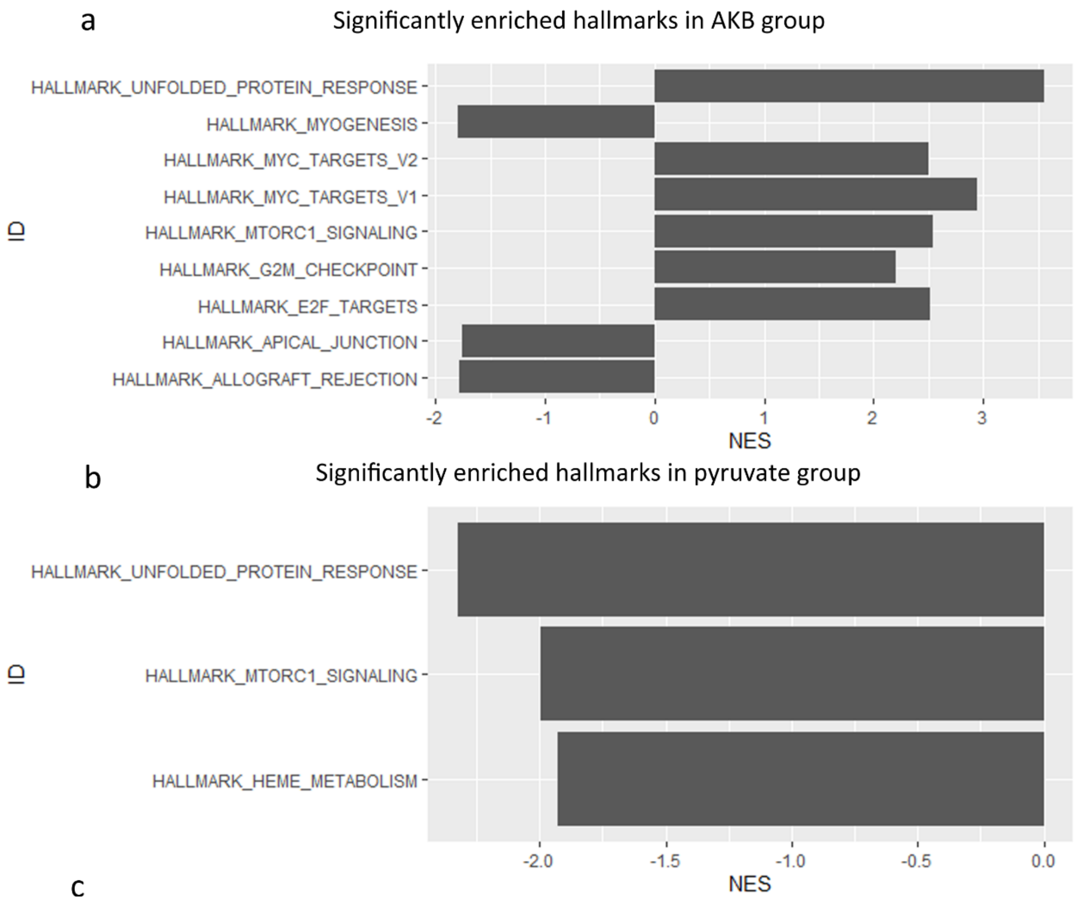


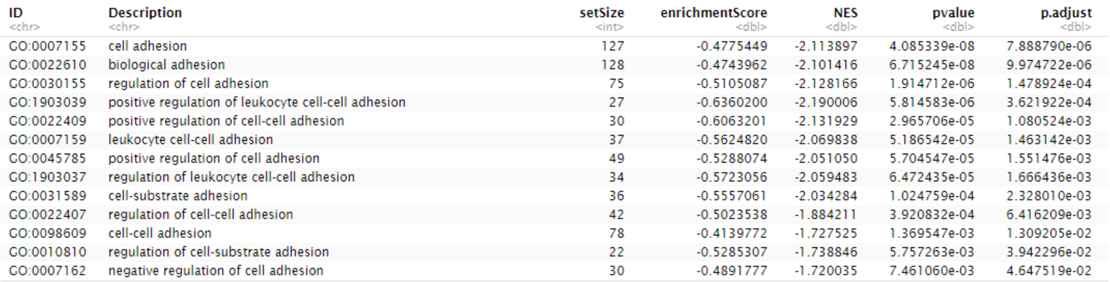
Figure 5. GO gene set enrichment analysis result. **a** GO gene set enrichment analysis for AKB group. **b** GO gene set enrichment analysis for pyruvate group.

GSEA and GO analysis show AKB induces cancer progression and metastasis

Before GSEA, all DEGs were decreasingly ranked with the log2 fold change, and all Ensembl IDs needs to be converted to Entrez IDs. Genes with Ensembl IDs which cannot be matched to Entrez IDs were discarded. Total 1640 (AKB group) and 543 (pyruvate group) genes were analyzed.



c



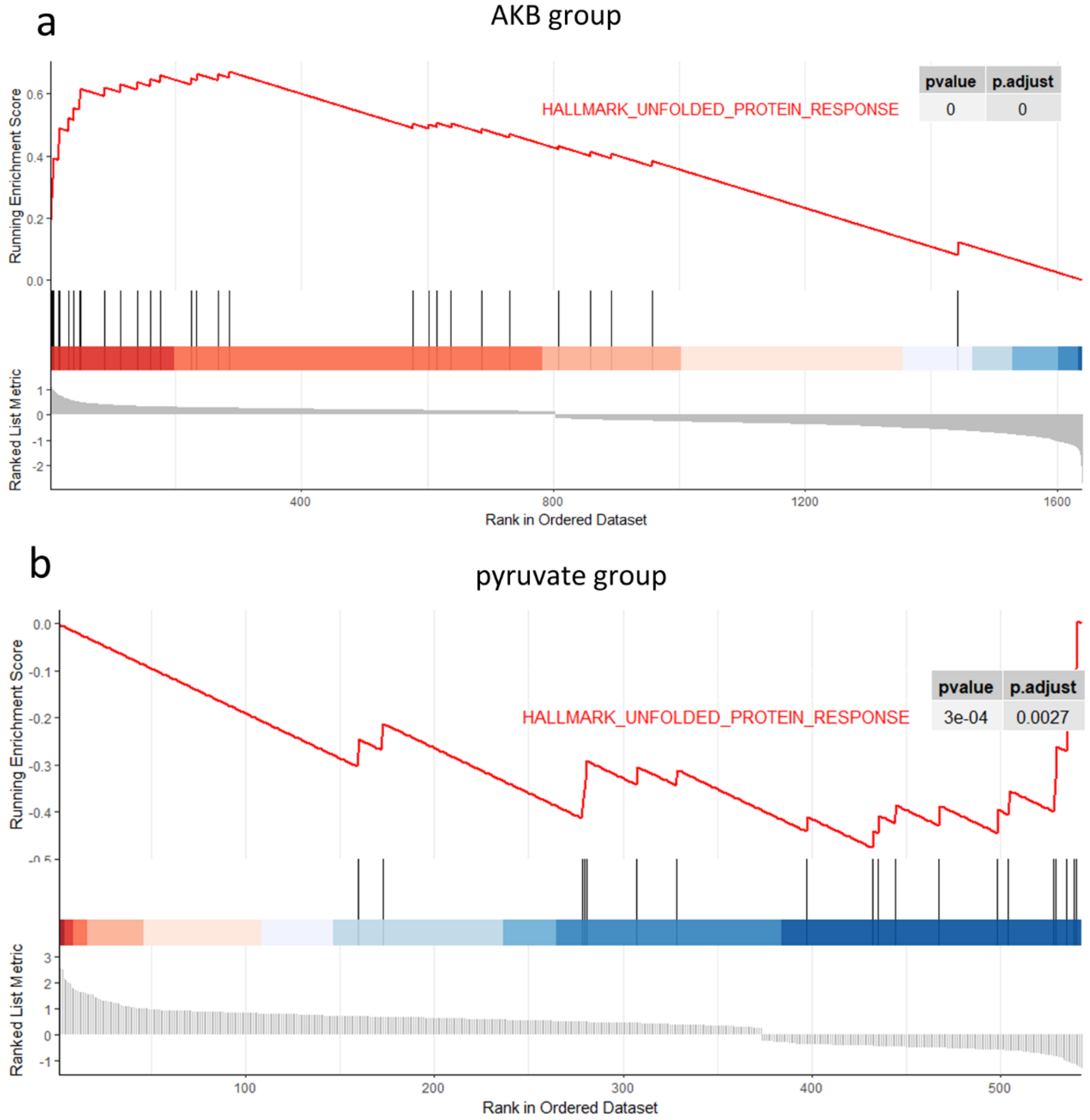
ID	Description	setSize	enrichmentScore	NES	pvalue	p.adjust
<chr>	<chr>	<int>	<dbl>	<dbl>	<dbl>	<dbl>
GO:0007155	cell adhesion	127	-0.4775449	-2.113897	4.085339e-08	7.888790e-06
GO:0022610	biological adhesion	128	-0.4743962	-2.101416	6.715245e-08	9.974722e-06
GO:0030155	regulation of cell adhesion	75	-0.5105087	-2.128166	1.914712e-06	1.478924e-04
GO:1903039	positive regulation of leukocyte cell-cell adhesion	27	-0.6360200	-2.190006	5.814583e-06	3.621922e-04
GO:0022409	positive regulation of cell-cell adhesion	30	-0.6063201	-2.131929	2.965706e-05	1.080524e-03
GO:0007159	leukocyte cell-cell adhesion	37	-0.5624820	-2.069838	5.186542e-05	1.463142e-03
GO:0045785	positive regulation of cell adhesion	49	-0.5288074	-2.051050	5.704547e-05	1.551476e-03
GO:1903037	regulation of leukocyte cell-cell adhesion	34	-0.5723056	-2.059483	6.472435e-05	1.666436e-03
GO:0031589	cell-substrate adhesion	36	-0.5557061	-2.034284	1.024759e-04	2.328010e-03
GO:0022407	regulation of cell-cell adhesion	42	-0.5023538	-1.884211	3.920832e-04	6.416209e-03
GO:0098609	cell-cell adhesion	78	-0.4139772	-1.727525	1.369547e-03	1.309205e-02
GO:0010810	regulation of cell-substrate adhesion	22	-0.5285307	-1.738846	5.757263e-03	3.942296e-02
GO:0007162	negative regulation of cell adhesion	30	-0.4891777	-1.720035	7.461060e-03	4.647519e-02

Figure 6. Hallmark GSEA and cell adhesion GO gene set enrichment analysis. **a-b** Enrichment of hallmarks in AKB and pyruvate group (all adjusted p-value < 0.01, not shown). **c** Enrichment of cell adhesion related GO terms in AKB group.

Using GSEA function and hallmark gene sets, several enriched hallmarks were identified in each group. All GSEAs shown in Figure 6a and 6b have adjusted p-value lower than 0.01. DEGs of AKB group are highly enriched in oncogene MYC targets hallmark with normalized enrichment score larger than 2.5. GSEA of apical junction shows that the cell-cell adhesion is downregulated after AKB treatment (Figure 6a). GO gene set enrichment analysis also shows a decrease in cell adhesion (Figure 6c). The upregulated oncogenes and downregulated cell adhesion suggest that AKB induces cancer progression and metastasis (Schulze et al., 2020). There is no obvious cancer progression hallmark enrichment in the pyruvate group (Figure 6b).

Different electron acceptors may induce opposite cellular changes

Unfolded protein response is a cellular stress response related to the endoplasmic reticulum. GSEA of the AKB and pyruvate group show the opposite results of unfolded protein response enrichment, which indicates that the different electron acceptors may have distinct effects on cellular stress (Figure 7).


AKB group

pyruvate group

Figure 7. Unfolded protein response hallmark GSEA plot of both groups

Discussion

In summary, I performed RNA-seq data analysis on electron acceptor treated K562 cells. Generally, AKB exerts a more potent effect on cellular changes than pyruvate. AKB treated K562 cells exhibit an increase in oncogene expression and decrease in cell-cell adhesion, which suggests that AKB promotes cancer progression and metastasis. AKB can also induce apoptotic process, which is possibly caused by oxidative stress. Through GO gene set enrichment analysis, I found that AKB induces enrichment in RNA processing and metabolic process, while in the pyruvate group, DEGs are enriched organic acid metabolic process and transmembrane transport. Moreover, different electron acceptors may exert opposite effects on some cellular processes like unfolded protein response. But the opposite effect may also be caused by the loss of some important DEGs during ID conversion, so it is not 100% certain that AKB and pyruvate did have the opposite effects.

Unexpectedly, only a very few DEGs with adjusted p-value lower than 0.05 and log2 fold change > 2 or < -2 were identified. Figure 8 is the heatmap of the top 20 variable genes in the AKB and pyruvate group. Visual inspection shows that the colour change (fold change) is not so distinct between the treated and control group and the replicates are also not clustered very well. Since the alignment rate and assignment rate (feature count step) are reasonable, it might be the problem with the wet lab experiments. The samples were treated with AKB or pyruvate for 24 hours. It may be too short to induce the transcriptional response since the modifications on transcriptional level are much slower than post-transcriptional modification. Moreover, the amount of AKB and pyruvate added may not be enough to exert distinct effects on the transcriptome.

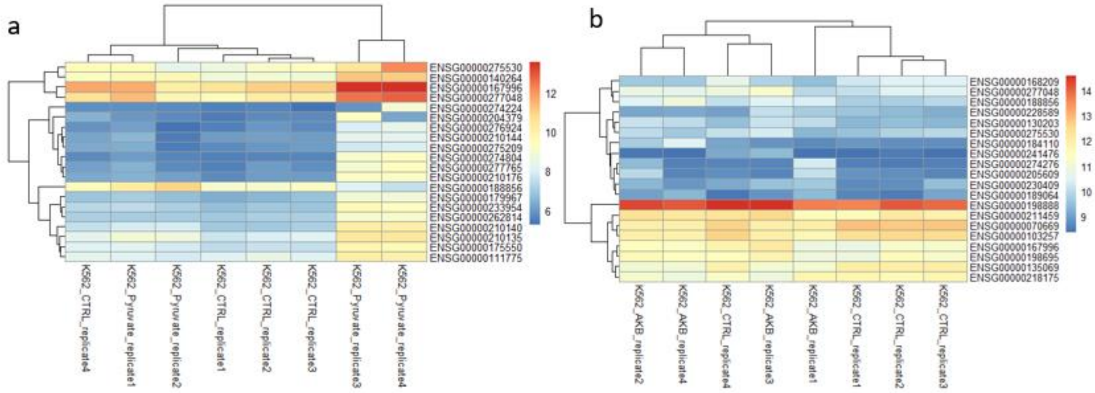
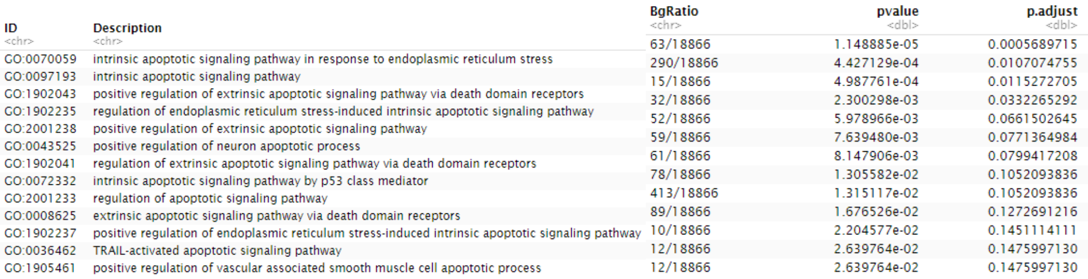


Figure 8. Heatmap of top 20 variable genes (use vst normalization) in AKB (a) and pyruvate (b) group

We should be careful when interpreting the results of apoptosis (Figure 4a) since only 4 genes have log2 fold change > 1 in the AKB group. I set the threshold of log2 fold change to 0.5 (total 49 genes) and found that there are still apoptotic signalling pathways significantly enriched, which confirms the apoptosis effect of AKB.



ID	Description	BgRatio	pvalue	p.adjust
<chr>	<chr>	<chr>	<dbl>	<dbl>
GO:0070059	intrinsic apoptotic signaling pathway in response to endoplasmic reticulum stress	63/18866	1.148885e-05	0.0005689715
GO:0097193	intrinsic apoptotic signaling pathway	290/18866	4.427129e-04	0.0107074755
GO:1902043	positive regulation of extrinsic apoptotic signaling pathway via death domain receptors	15/18866	4.987761e-04	0.0115272705
GO:1902235	regulation of endoplasmic reticulum stress-induced intrinsic apoptotic signaling pathway	32/18866	2.300298e-03	0.0332265292
GO:2001238	positive regulation of extrinsic apoptotic signaling pathway	52/18866	5.978966e-03	0.0661502645
GO:0043525	positive regulation of neuron apoptotic process	59/18866	7.639480e-03	0.0771364984
GO:1902041	regulation of extrinsic apoptotic signaling pathway via death domain receptors	61/18866	8.147906e-03	0.0799417208
GO:0072332	intrinsic apoptotic signaling pathway by p53 class mediator	78/18866	1.305582e-02	0.1052093836
GO:2001233	regulation of apoptotic signaling pathway	413/18866	1.315117e-02	0.1052093836
GO:0008625	extrinsic apoptotic signaling pathway via death domain receptors	89/18866	1.676526e-02	0.1272691216
GO:1902237	positive regulation of endoplasmic reticulum stress-induced intrinsic apoptotic signaling pathway	10/18866	2.204577e-02	0.1451114111
GO:0036462	TRAIL-activated apoptotic signaling pathway	12/18866	2.639764e-02	0.1475997130
GO:1905461	positive regulation of vascular associated smooth muscle cell apoptotic process	12/18866	2.639764e-02	0.1475997130

(Since the table is too wide, it has to be broken into two parts.)

Another limitation is that GSEA was only performed on part of the gene sets, there might be some other gene sets that I did not notice but their GSEA results may contradict my conclusion. Focusing on only some gene sets may introduce bias to the analysis. However, it is also difficult to integrate the information of all highly enriched gene sets. Lots of gene set enrichment analysis results can be quite chaotic and difficult to interpret. Moreover, it is not possible to use some gene sets since they are produced in a specific context, for example with a particular gene knockout. Since I am not addressing a very specific question, those gene sets may not be helpful.

References

GORRINI, C., HARRIS, I. S. AND MAK, T. W. (2013) 'Modulation of oxidative stress as an anticancer strategy', *Nature Reviews Drug Discovery*, 12(12), pp. 931–947. doi: 10.1038/nrd4002.

KOEFFLER, H. P. AND GOLDE, D. W. (1980) 'Human myeloid leukemia cell lines: a review.', *Blood*. United States, 56(3), pp. 344–350.

LIAO, Y., SMYTH, G. K. AND SHI, W. (2014) 'featureCounts: an efficient general purpose program for assigning sequence reads to genomic features', *Bioinformatics*, 30(7), pp. 923–930. doi: 10.1093/bioinformatics/btt656.

PERTEA, M., KIM, D., PERTEA, G. M., LEEK, J. T. AND SALZBERG, S. L. (2016) 'Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown', *Nature Protocols*, 11(9), pp. 1650–1667. doi: 10.1038/nprot.2016.095.

SCHULZE, A., OSHI, M., ENDO, I. AND TAKABE, K. (2020) 'MYC Targets Scores Are Associated with Cancer Aggressiveness and Poor Survival in ER-Positive Primary and Metastatic Breast Cancer', *International Journal of Molecular Sciences* . doi: 10.3390/ijms21218127.

SCHURCH, N. J., SCHOFIELD, P., GIERLIŃSKI, M., COLE, C., SHERSTNEV, A., SINGH, V., WROBEL, N., GHARBI, K., SIMPSON, G. G., OWEN-HUGHES, T., BLAXTER, M. AND BARTON, G. J. (2016) 'How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?', *RNA (New York, N.Y.)*, 22(6), pp. 839–851. doi: 10.1261/rna.053959.115.

YU, G., WANG, L.-G., HAN, Y. AND HE, Q.-Y. (2012) 'clusterProfiler: an R package for comparing biological themes among gene clusters.', *Omics : a journal of integrative biology*, 16(5), pp. 284–287. doi: 10.1089/omi.2011.0118.

Supplementary information

Table S1. Software version

Software/R package name	Version
FastQC	0.11.9
Cutadapt	3.2 with Python 3.6.6
HISAT2	2.1.0
featureCounts	2.0.1
DESeq2	1.30.1
clusterProfiler	3.18.1
org.Hs.eg.db	3.12.0

Table S2. Mapping statistics

Sample	**Number of Reads**	**aligned concordantly 0 times**	**aligned concordantly exactly 1 time**	**aligned concordantly >1 times**	**Overall alignment rate**
CTRL_replicate1	21696493	2108325 (9.72%)	18782094 (86.57%)	806074 (3.72%)	97.31%
CTRL_replicate2	19538109	1807997 (9.25%)	17001508 (87.02%)	728604 (3.73%)	97.64%

CTRL_replicate3	22042584	1929159 (8.75%)	19291114 (87.52%)	822311 (3.73%)	97.55%
CTRL_replicate4	19883196	1967573 (9.90%)	17166314 (86.34%)	749309 (3.77%)	97.36%
AKB_replicate1	18458174	1699878 (9.21%)	16086647 (87.15%)	671649 (3.64%)	97.42%
AKB_replicate2	19428844	1878867 (9.67%)	16812627 (86.53%)	737350 (3.80%)	97.19%
AKB_replicate3	20410792	1823952 (8.94%)	17832300 (87.37%)	754540 (3.70%)	96.95%
AKB_replicate4	19418134	1853731 (9.55%)	16834282 (86.69%)	730121 (3.76%)	97.06%
Pyruvate_replicate1	20579457	1781556 (8.66%)	18028236 (87.60%)	769665 (3.74%)	97.15%
Pyruvate_replicate2	20129938	2072275 (10.29%)	17321652 (86.05%)	736011 (3.66%)	97.10%
Pyruvate_replicate3	23411722	1555108 (6.64%)	20889386 (89.23%)	967228 (4.13%)	97.40%
Pyruvate_replicate4	25378222	1735523 (6.84%)	22563014 (88.91%)	1079685 (4.25%)	97.31%

Table S3. feature count statistics

Sample	CTRL_replicate1	CTRL_replicate2	CTRL_replicate3	CTRL_replicate4
Rate	71.3%	71.9%	71.6%	71.8%
Sample	AKB_replicate1	AKB_replicate2	AKB_replicate3	AKB_replicate4
Rate	71.5%	70.6%	71.8%	71.0%
Sample	Pyruvate_replicate1	Pyruvate_replicate2	Pyruvate_replicate3	Pyruvate_replicate4
Rate	71.4%	72.0%	72.2%	71.0%

Table S4. The source of the reference genome, gene annotation files and gene sets

Name	Extension	FTP/website
Reference genome	fasta	ftp://ftp.ensembl.org/pub/release-84/fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna.primary_assembly.fa.gz
Gene annotation	gtf	ftp://ftp.ensembl.org/pub/release-84/gtf/homo_sapiens/Homo_sapiens.GRCh38.84.gtf.gz
Gene sets	gmt	https://www.gsea-msigdb.org/gsea/downloads.jsp#msigdb

Table S5. Command lines

Software	Usage	Command line

FastQC	Quality control	fastqc input.fq.gz
Cutadapt	Trim the first 14 bases	cutadapt -u 14 -o output.fq.gz input.fq.gz
HISAT2	Create index	hisat2-build -p 4 genome.fa genome
	Alignment for pair-end data	hisat2 -p 4 -x reference_index -S output.sam -1 input_1.fq.gz -2 input_2. fq.gz 2>aligned_info.txt
featureCounts	Assign sequence reads to genomic features	featureCounts -T 4 -t exon -g gene_id -p -a input.gtf -o all_sam_files.sam

Code and data availability

All R codes, read count matrix, DEG data and other files have been deposited in the following GitHub repository:

https://github.com/Leo010530/GP2_Data_Analysis_ICA