The histone demethylase JMJD2B regulates endothelial-to-mesenchymal transition

Wong Jun Hong A0182774U
Sheryl Poh Jing Wen A0189139N

Introduction

Endothelial cells (**ECs**) play an important role in maintenance of the vascular system and the repair after injury. Under proinflammatory conditions, endothelial cells can acquire a mesenchymal phenotype by a process named endothelial-to-mesenchymal transition (**EndMT**), which affects the functional properties of endothelial cells (J C. Kovacic, MD, PhD, 2019). From past research, it is understood that by inhibiting a specific histone demethylase, JMJD2B, EndMT is reduced (S Glaser et al., 2020). However, the proteins responsible and the pathways influenced were not shown.

In this study, we will identify the proteins contributing to the decrease of EndMT during JMJD2B inhibition. We exposed human ECs of two types: one Wild-type (**WT**) and one with an siRNA knock-down of JMJD2B (**KD**), to two media: Differential Medium (**DIFF**), which promotes EndMT and Full medium (**FULL**), which will be the control. Microarray expression data is analysed, and the programming language R is used to carry out data manipulation to identify features meeting our specified cutoffs, and subsequently to map transcript clusters to their gene names and gene IDs. REACTOME is then used to map the gene IDs of interest to their respective biological pathways. Genes encoding proteins that reduce EndMT during JMJD2B inhibition are identified using ENSEMBL. Line plots will be used to show expression levels of these genes in varying conditions. Using knowledge of proteins that are known to induce EndMT from previous reports as indicators, we identified genes encoding for proteins that reduce EndMT during JMJD2B inhibition. Possible protein mechanisms reducing EndMT were also identified.

Methods and Materials

Gene Expression Omnibus (GEO) Dataset

GEO provides us with the expression data needed to observe differential gene expression. Here, we used the expression data of genes from Affymetrix Human Exon 1.0 ST Array, in full and differentiation medium under control and JMJD2B knockdown.

RStudio and REACTOME

RStudio is an integrated development environment for R, a programming language for statistical computing and graphics. We used the packages **GEOquery**, **affy**, **limma** and **oligo** from **Bioconductor** for oligonucleotide array analysis. Next, **huex10sttranscriptcluster.db** is used for mapping transcript IDs to their gene names and Entrez gene ID. We used **reactomePA** and **reactome.db** from REACTOME when mapping Entrez gene IDs to pathways. To present our data, we used **ggplot2, readr, ggpubr, and formattable**. The script is attached to the Annex.

ENSEMBL Genome Browser

Information such as gene sequence, splice variants and further annotation can be retrieved at the genome, gene and protein level using ENSEMBL. Here, we used ENSEMBL to identify the genes encoding for proteins that inhibit Interleukin-1β and TGFβ.

Interleukin-1β and TGFβ

Interleukin-1β and TGFβ are known to induce EndMT (Seol, M.A., Kim, J., Oh, K. et al., 2019). Genes encoding for proteins that inhibits the pathways of either Interleukin-1β and TGFβ will be classified as EndMT-reducing.

Identifying proteins reducing EndMT during JMJD2B inhibition

The Microarray expression data was obtained from GEO (**GSE143150**) which comprises 12 gene expression files (Fig. 1). These file samples are categorised into 4 different conditions (Media/Genotype): Differentiation Media/Wild-type (DIFF_WT), Differentiation Media/Knockdown (DIFF_KD), Full media/Wild-type (FULL_WT) and Full media/Knockdown (FULL_KD). We mainly focused on the expression data of every transcript cluster.

	treatment	genotype	cond
GSM4250986	treatment: differentiation media (DM)	genotype/variation: Wild-type	DIFF_WT
GSM4250987	treatment: differentiation media (DM)	genotype/variation: JMJD2B knockdown	DIFF_KD
GSM4250988	treatment: full medium (FM)	genotype/variation: Wild-type	FULL_WT
GSM4250989	treatment: full medium (FM)	genotype/variation: JMJD2B knockdown	FULL_KD
GSM4250990	treatment: differentiation media (DM)	genotype/variation: Wild-type	DIFF_WT
GSM4250991	treatment: differentiation media (DM)	genotype/variation: JMJD2B knockdown	DIFF_KD
GSM4250992	treatment: full medium (FM)	genotype/variation: Wild-type	FULL_WT
GSM4250993	treatment: full medium (FM)	genotype/variation: JMJD2B knockdown	FULL_KD
GSM4250994	treatment: differentiation media (DM)	genotype/variation: Wild-type	DIFF_WT
GSM4250995	treatment: differentiation media (DM)	genotype/variation: JMJD2B knockdown	DIFF_KD
GSM4250996	treatment: full medium (FM)	genotype/variation: Wild-type	FULL_WT
GSM4250997	treatment: full medium (FM)	genotype/variation: JMJD2B knockdown	FULL_KD

Fig. 1: Summary of the cell cultures under different conditions used in the experiment

Background correction, normalisation and expression calculation is done using the function rma(). Due to the multifactorial character of the GEO dataset, we had to use the Empirical Bayes method on the dataset (method eBayes in R). The steps taken were as follows:

```
expression_data <- oligo::rma(data) # Background correction, Normalistation using rma() on dataset
eset <- exprs(expression_data)
model <- model.matrix( ~ 0 + expression_data$cond) #linear model, with intercept and the coefficient for all conditions ("DIFF_KD","DIFF_WT","FULL_KD","FULL_WT")
colnames(model) <- levels(expression_data$cond)
contrasts <- makeContrasts(DIFF_KD - DIFF_WT, #Contrast between genotypes(KD and WT) in DIFF medium
                           FULL_KD - FULL_WT, #Contrast between genotypes(KD and WT) in FULL medium
                           FULL_KD - DIFF_KD, #Contrast between media(DIFF and FULL) in KD genotype
                           FULL_WT - DIFF_WT, #Contrast between media(DIFF and FULL) in WT genotype
                           interaction=(DIFF_KD-DIFF_WT) - (FULL_KD - FULL_WT), #Contrast between different genotype with different media, also known as interaction
                           (DIFF_KD - DIFF_WT) + ( FULL_KD - FULL_WT), #Contrast between genotypes across media
                           (FULL_KD - DIFF_KD) + ( FULL_WT - DIFF_WT), #Contrast between media across genotypes
                           levels = model)

expdata_fitted_contrasts <- lmFit(expression_data,model) #Expression data undergoes Empirical Bayes method w.r.t linear model
fitted.contrasts <- contrasts.fit(expdata_fitted_contrasts,contrasts) #Subsequently fitted with the 7 differnt contrasts above
fitted.aebayes <- eBayes(fitted.contrasts) #Dataset with Empirical Bayes method applied
```

P-value of 0.05 and Log Fold Change (logFC) value of 1 was set as the threshold for the genes. We then tabulated which contrasts hold most of our transcript clusters passing cutoff (Fig. 2). We also narrowed our attention to the contrast between different media in KD (**Contrast 1**), the contrast between different media in WT (**Contrast 2**), and the contrast between different media across both genotypes (**Contrast 3**). Volcano plots were used to depict the Log fold change in expression levels of the genes in Contrasts 1,2 and 3, against their p-values (Fig. 3a, 3b and 3c).

	Genotype_in_DIFF	Genotype_in_FULL	Media_in_KD	Media_in_WT	Assuming_Interaction	Genotypes_across_media	Media_across_genotypes
Transcript clusters with p-values<0.05 and logFC>1	0	0	194	89	0	0	576

Fig. 2: Number of transcripts passing cutoff (LogFC >1 and p-value < 0.05) under each contrast

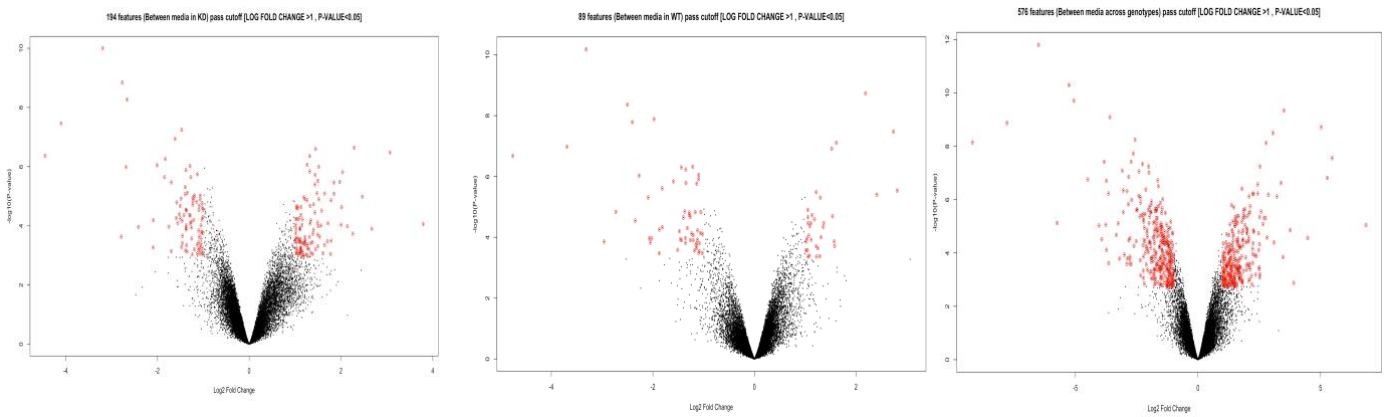

Fig. 3a,3b and 3c: Plotted volcano plots for transcript clusters present in our dataset. The points circled in red shows the features that meet the criterion for logFC>1 and p-value<0.05

Transcript clusters meeting the specified cutoff from Contrasts 1,2 and 3 were mapped to their Entrez gene IDs. Mapped transcript clusters of Contrast 1 (**MC1**), Contrast 2 (**MC2**) and Contrast 3 (**MC3**) are obtained. Finding genes that are uniquely found in MC1 but not in MC3 allows us to filter out the differentially expressed genes specific to the ECs of KD genotype across media (**Genes_KD**). Similarly, genes uniquely found in MC2 but not in MC3 are the differentially expressed genes specific to ECs of the WT genotype across media (**Genes_WT**). There were a total of 20 genes found in Genes_KD and only 2 genes in Genes_WT (Figure 4a and 4b).

transcript_cluster_id	SYMBOL	GENENAME	ENSEMBLID	ENTREZID
2592005	HIBCH	3-hydroxyisobutyryl-CoA hydrolase	ENSG00000198130	26275
3595846	MINDY2	MINDY lysine 48 deubiquitinase 2	ENSG00000128923	54629
2550790	LRPPRC	leucine rich pentatricopeptide repeat containing	ENSG00000138095	10128
2520533	NABP1	nucleic acid binding protein 1	ENSG00000173559	64859
2785282	SCLT1	sodium channel and clathrin linker 1	ENSG00000151466	132320
2813060	PIK3R1	phosphoinositide-3-kinase regulatory subunit 1	ENSG00000145675	5295
2343823	ADGRL2	adhesion G protein-coupled receptor L2	ENSG00000117114	23266
2789266	LRBA	LPS responsive beige-like anchor protein	ENSG00000198589	987
2967276	POPDC3	popeye domain containing 3	ENSG00000132429	64208
2536965	LOC285097	uncharacterized FLJ38379	ENSG00000280119	285097
2468920	CPSF3	cleavage and polyadenylation specific factor 3	ENSG00000119203	51692
2577958	DARS1	aspartyl-tRNA synthetase 1	ENSG00000115866	1615
2406420	CLSPN	claspin	ENSG00000092853	63967
2781736	CFI	complement factor I	ENSG00000205403	3426
2929036	LTV1	LTV1 ribosome biogenesis factor	ENSG00000135521	84946
2639054	PARP14	poly(ADP-ribose) polymerase family member 14	ENSG00000173193	54625
2441386	RGS5	regulator of G protein signaling 5	ENSG00000143248,ENSG00000232995	8490
2914777	TTK	TTK protein kinase	ENSG00000112742	7272
2428796	PTPN22	protein tyrosine phosphatase non-receptor type 22	ENSG00000134242	26191
2555490	XPO1	exportin 1	ENSG00000082898	7514

Fig.4a: 20 differentially expressed genes found in KD across media.

transcript_cluster_id	SYMBOL	GENENAME	ENSEMBLID	ENTREZID
3147985	LRP12	LDL receptor related protein 12	ENSG00000147650	29967
2900059	H2BC14	H2B clustered histone 14	ENSG00000273703	8342

Fig.4b: 2 differentially expressed genes found in WT across media.

To find all the pathways affected when JMJD2B is inhibited, we parsed Gene_KD through REACTOME. We managed to locate 162 of such pathways (Fig 5).

pathways affected in KD cells across media (EndMT transition)
Extra-nuclear estrogen signaling
TCR signaling
PI3K events in ERBB4 signaling
Signaling by FGFR3 fusions in cancer
Signaling by FGFR4 in disease
GP1b-IX-V activation signalling
Erythropoietin activates Phosphoinositide-3-kinase (PI3K)
Constitutive Signaling by EGFRvIII
Signaling by EGFRvIII in Cancer
PI3K events in ERBB2 signaling

Fig.5: 10 (of 162) pathways identified from Genes_KD, in no particular order. Remaining data is attached to Annex 1.

Line plots were used to observe expression levels of genes in X in different conditions. Expression values of each gene were obtained by using the expression data after background correction and normalisation. After which, we used the mean expression values from all 4 different conditions of each gene (Fig.6).

EndMT is induced by proteins such as Interleukin-1β and TGFβ (Jin Gu Cho, Aram Lee,Woochul Chang,Myeong-Sok Lee, Jongmin Kim, 2018). Some of the genes found in Gene_KD encodes for proteins that are related to the inhibition of those EndMT-inducing proteins. Using ENSEMBL, we identified 3 genes from Gene_KD that encode for these proteins and classified them as **X.** We also found that the 2 genes from Gene_WT but they were not EndMT-related. Hence, we decided to focus only on Gene_KD.

Using data from Fig.5, we identified the possible ways gene products from X could have reduced EndMT in KD ECs (Fig.7). This was done by mapping the pathways each gene in X was responsible for.

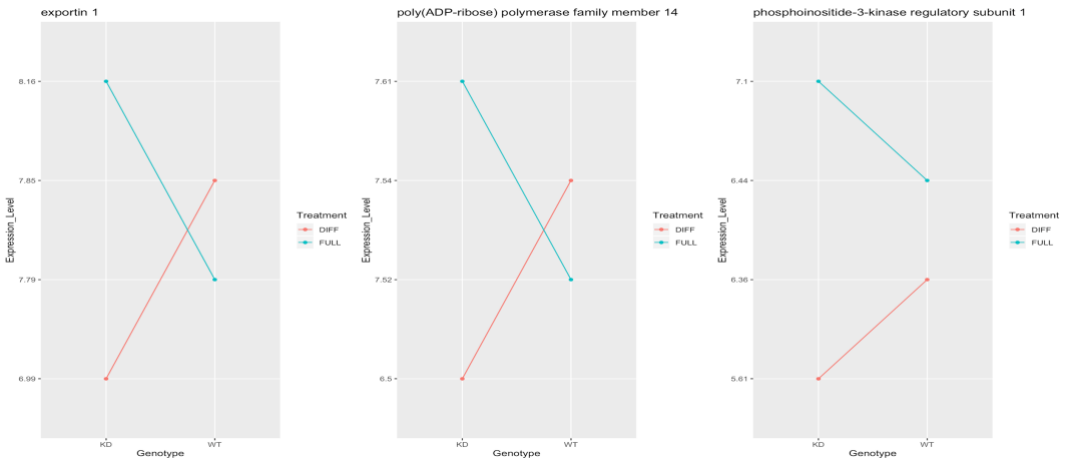

Fig.6: Expression levels of 3 classified genes in X in 4 different conditions: DIFF_WT, DIFF_KD, FULL_WT, FULL_KD.

Genes from X	Pathways reducing EndMT during JMJD2B Knockdown
Exportin 1	- Downregulation of TGF-beta receptor signaling - TGF-beta receptor signaling activates SMADs - Signaling by TGF-beta Receptor Complex
Poly(ADP-ribose) polymerase family member 14	- Metabolism of water-soluble vitamins and cofactors - Metabolism of vitamins and cofactors - Nicotinamide salvaging - Nicotinate metabolism
Phosphoinositide-3-kinase regulatory subunit 1	- Cell-Cell communication - Cell surface interactions at the vascular wall pathways - Interleukin receptor SHC signaling - Interleukin-7 signaling - Interleukin-2 family signaling - Interleukin-3, Interleukin-5 and GM-CSF signaling - Interleukin-4 and Interleukin-13 signaling

Fig.7: Possible ways X could have reduced EndMT during JMJD2B knockdown.

DISCUSSION AND ANALYSIS OF RESULTS

With the affy package in R, we utilised the RMA method to do background correction and normalisation of the datasets analysed. Background correction is done to eliminate background noise that arises non-specific hybridisation, overshining or technical imperfections (Sifakis, 2012). Furthermore, normalisation is done to the corrected data that are affected by experimental inconsistencies such as limited sampling, differences in array production batches, hybridization and washing conditions, scanning power, etc (Terri T Ni, 2008). A simple eBayes() function then uses the empirical Bayes method to shrink the individual probe-wise sample variances towards a common value that represents the overall distribution.

To determine if a gene is considered differentially expressed in a cell, we narrow down our data to look at genes that are expressed differently by a worthwhile amount using a fold change of 1. However fold-change cutoffs do not take into account reliability and reproducibility of the result. Therefore, it is important to also ensure that our data satisfies the p-value criteria of less than 0.05. This means that there is only a 5% chance of obtaining a false positive (McCarthy, 2009).

Contrasts were made between datasets that were of the KD/WT genotype or in the FULL/DIFF media to further study the effects of media change and knockdown of JMJD2B activity on the ECs. Identification of the statistically significant genes revealed that the differentially expressed genes were caused by the change in media in both WT and KD genotypes. Volcano plots of this data were then plotted as shown in Fig. 3a,b and c.

After filtering out our data to focus on the effects of media in the ECs of WT and KD genotype, we identified 22 genes that were differentially expressed between different media. Of them consists of 20 genes from the KD cell and 2 from the WT cell. We then decided to take a closer look into the biological pathway these proteins are involved in through the reactome platform.

In our study, we also used proteins such as Interleukin-1β and TGFβ as indicators of EndMT. Using these protein indicators, we picked candidates out of the 22 genes to be classified as genes that contribute to reducing EndMT in JMJD2B KD ECs.

Using ENSEMBL, we identified 3 of such genes from the KD cell. The other 2 genes from WT cells encodes for proteins that mostly regulates cell metabolism during inflammatory response, and not directly affecting EndMT. Expression level of these 3 genes from KD ECs decreased in differential medium, as shown in Fig.6. As differential medium simulates EndMT, these 3 genes were downregulated in EndMT when JMJD2B is inhibited. In full medium, which is not EndMT-inducing, these 3 genes were upregulated when JMJD2B is inhibited. Reasons as to why these genes were upregulated or downregulated are to be addressed in future research.

We then evaluated the proteins encoded by the three genes related to the decrease in EndMT based on the mechanisms that are unique to the cells that had an inhibition of the histone demethylase JMJD2B.

Exportin 1

The exportin 1 protein is involved in many pathways related to the EndMT process. When parsed through REACTOME, we got pathways such as TGF-beta receptor signaling activates SMADs and signaling by TGF-beta Receptor Complex. TGF-beta is a dimeric cytokine produced from various cells in an inactive form. After activated through cleavage, it sends signals to its receptors when in turn phosphorylates and activates SMAD pathways (Pardali,2017). SMAD pathways are upregulated and forms SMAD complexes that can act as transcriptional activators that increase expression of mesenchymal markers such as alpha smooth muscle actin (SMA) which then leads to increased EndMT (Jin Gu Cho, Aram Lee,Woochul Chang,Myeong-Sok Lee, Jongmin Kim, 2018). An reduced expression of exportin 1 found in cells of the KD genotype shows that there is a downregulation of TGF-beta receptor signalling.This contributes to a lower expression of mesenchymal genes and therefore decreased EndMT, which shows that the inhibition of JMJD2B using siRNA indeed affected the EndMT process.

Phosphoinositide-3-kinase regulatory subunit 1 (PIK3R1)

PIK3R1 is also found to contribute to the EndMT process. PIK3R1 is involved in many pathways related to interleukin signalling including Interleukin receptor SHC signaling, and signalling of Interleukin 2,3,4,5,7,13 as seen from Fig.7. Interleukin is a well-known inducer of the EndMT process. Interleukin 7 for example, when, used in treatment for cells increased the transcription of EndMT-related genes (Seol, M.A., Kim, J., Oh, K. et al., 2019).

Furthermore, PIK3R1 plays a part in Cell-Cell communication as well as cell surface interactions at the vascular wall pathways. EndMT is a process marked by a decrease in intercellular adhesion forces in monolayer and cell stiffening and flattening (Ana Sancho, Vandersmissen,Sander Craps,Aernout Luttun,and Jürgen Grollb, 2017). The downregulation of PIK3R1 in KD cells when exposed to hypoxic conditions can suggest that there was limited modulation of cell to cell communication as well as cell surface interactions, which made EndMT less likely to occur. The decreased level of interleukin signalling also inhibits transcription of mesenchymal genes.

Poly(ADP-ribose) polymerase family member 14 (PARP-14)

PARP-14 is a member of the poly(ADP-ribose) polymerase family. Other than being involved in metabolism and cell death, PARP-14 may also induce inflammatory responses by promoting gene expression of related genes, including interleukin (IL)-1β, tumor necrosis factor (TNF)-α and endothelin-1 (Yan, F., Zhang, G., Feng, M. et al., 2015). These genes combined lead to a heightened expression of EndMT-related genes. When PARP-14 is expressed at a lower level, it therefore leads to less endothelial to mesenchymal transitions.

Conclusion

From our results, we found 3 genes that contributed to reducing EndMT during JMJD2B inhibition. We also found possible ways these 3 genes could have reduced EndMT. Further research can be conducted to understand how the regulation of these genes influenced their respective mechanisms.

References

J C. Kovacic, MD, PhD. 2019. Endothelial to Mesenchymal Transition in Cardiovascular Disease. JACC STATE-OF-THE-ART REVIEW: http://www.onlinejacc.org/content/73/2/190

Glaser, S. F., Heumüller, A. W., Tombor, L., Hofmann, P., Muhly-Reinholz, M., Fischer, A., … Dimmeler, S. (2020, February 25). The histone demethylase JMJD2B regulates endothelial-to-mesenchymal transition. Retrieved from https://www.pnas.org/content/117/8/4180

Sifakis, E. G., Prentza, A., Koutsouris, D., & Chatziioannou, A. A. (2011, November 8). Evaluating the effect of various background correction methods regarding noise reduction, in two-channel microarray data. Retrieved from https://www.sciencedirect.com/science/article/pii/S0010482511001995

Ni, T. T., Lemon, W. J., Shyr, Y., & Zhong, T. P. (2008, November 28). Use of normalization methods for analysis of microarrays containing a high degree of gene effects. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2612699/

Davis J. McCarthy, Gordon K. Smyth, Testing significance relative to a fold-change threshold is a TREAT, Bioinformatics, Volume 25, Issue 6, 15 March 2009, Pages 765–771, https://doi.org/10.1093/bioinformatics/btp053

Pardali, E., Sanchez-Duffhues, G., Gomez-Puerto, M. C., & Ten Dijke, P. (2017, October 17). TGF-β-Induced Endothelial-Mesenchymal Transition in Fibrotic Diseases. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/29039786

Cho, J. G., Lee, A., Chang, W., Lee, M.-S., & Kim, J. (2018, February 20). Endothelial to Mesenchymal Transition Represents a Key Link in the Interaction between Inflammation and Endothelial Dysfunction. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/29515588

Sancho, A., Vandersmissen, I., Craps, S. et al. A new strategy to measure intercellular adhesion forces in mature cell-cell contacts. Sci Rep 7, 46152 (2017). https://doi.org/10.1038/srep46152

Seol, M.A., Kim, J., Oh, K. et al. Interleukin-7 Contributes to the Invasiveness of Prostate Cancer Cells by Promoting Epithelial–Mesenchymal Transition. Sci Rep 9, 6917 (2019). https://doi.org/10.1038/s41598-019-43294-4

Yan, F., Zhang, G., Feng, M. et al. Glucagon-Like Peptide 1 Protects against Hyperglycemic-Induced Endothelial-to-Mesenchymal Transition and Improves Myocardial Dysfunction by Suppressing Poly(ADP-Ribose) Polymerase 1 Activity. Mol Med 21, 15–25 (2015). https://doi.org/10.2119/molmed.2014.00259

Annex

162 genes from Figure 5

	pathways affected in KD cells
1	Extra-nuclear estrogen signaling
2	TCR signaling
3	PI3K events in ERBB4 signaling
4	Signaling by FGFR3 fusions in cancer
5	Signaling by FGFR4 in disease
6	GP1b-IX-V activation signalling
7	Erythropoietin activates Phosphoinositide-3-kinase (PI3K)
8	Constitutive Signaling by EGFRvIII
9	Signaling by EGFRvIII in Cancer
10	PI3K events in ERBB2 signaling
11	Role of LAT2/NTAL/LAB on calcium mobilization
12	GAB1 signalosome
13	Signaling by NTRK3 (TRKC)
14	Signaling by cytosolic FGFR1 fusion mutants
15	Tie2 Signaling
16	PI-3K cascade:FGFR3
17	Constitutive Signaling by Ligand-Responsive EGFR Cancer Variants
18	Nicotinamide salvaging
19	Translocation of ZAP-70 to Immunological synapse
20	Signaling by Ligand-Responsive EGFR Variants in Cancer
21	Branched-chain amino acid catabolism
22	Processing of Intronless Pre-mRNAs
23	PI-3K cascade:FGFR4
24	PI-3K cascade:FGFR1
25	Phosphorylation of CD3 and TCR zeta chains
26	CD28 dependent PI3K/Akt signaling
27	Signaling by FGFR3 in disease
28	Signaling by FGFR3 point mutants in cancer
29	Regulation of signaling by CBL
30	Nephrin family interactions
31	PI-3K cascade:FGFR2
32	Cytosolic tRNA aminoacylation
33	Estrogen-dependent nuclear events downstream of ESR-membrane signaling
34	G alpha (q) signalling events
35	Signaling by EGFR in Cancer

36	Role of phospholipids in phagocytosis
37	Downstream signaling of activated FGFR3
38	Cyclin A/B1/B2 associated events during G2/M transition
39	Signaling by NTRK2 (TRKB)
40	Signaling by Erythropoietin
41	ESR-mediated signaling
42	Downregulation of TGF-beta receptor signaling
43	Downstream signaling of activated FGFR4
44	Interleukin receptor SHC signaling
45	Processing of Capped Intronless Pre-mRNA
46	Downstream signal transduction
47	DAP12 signaling
48	Downstream signaling of activated FGFR2
49	NEP/NS2 Interacts with the Cellular Export Machinery
50	FGFR1 mutant receptor activation
51	Nicotinate metabolism
52	Downstream signaling of activated FGFR1
53	TGF-beta receptor signaling activates SMADs
54	Antigen activates B Cell Receptor (BCR) leading to generation of second messengers
55	Export of Viral Ribonucleoproteins from Nucleus
56	CD28 co-stimulation
57	Rev-mediated nuclear export of HIV RNA
58	GPVI-mediated activation cascade
59	Interleukin-7 signaling
60	Interactions of Rev with host cellular proteins
61	Activation of ATR in response to replication stress
62	Apoptotic cleavage of cellular proteins
63	Signaling by FGFR1 in disease
64	Signaling by FGFR3
65	Transport of Mature mRNA Derived from an Intronless Transcript
66	Signaling by FGFR4
67	RET signaling
68	Transport of Mature mRNAs Derived from Intronless Transcripts
69	Deactivation of the beta-catenin transactivating complex
70	tRNA Aminoacylation

71	Signaling by SCF-KIT
72	DAP12 interactions
73	Signaling by FGFR2 in disease
74	Cell Cycle Checkpoints
75	PI3K Cascade
76	Interleukin-2 family signaling
77	Signaling by Nuclear Receptors
78	Regulation of Complement cascade
79	IRS-mediated signalling
80	Interleukin-3, Interleukin-5 and GM-CSF signaling
81	Signaling by EGFR
82	Signaling by ERBB2
83	Signaling by FGFR1
84	IRS-related events triggered by IGF1R
85	Apoptotic execution phase
86	Synthesis of PIPs at the plasma membrane
87	IGF1R signaling cascade
88	Signaling by Type 1 Insulin-like Growth Factor 1 Receptor (IGF1R)
89	Insulin receptor signalling cascade
90	Signaling by ERBB4
91	Complement cascade
92	Signaling by PDGF
93	mRNA 3'-end processing
94	Signaling by FGFR in disease
95	Metabolism of amino acids and derivatives
96	RNA Polymerase II Transcription Termination
97	Costimulation by the CD28 family
98	Signaling by TGF-beta Receptor Complex
99	Signaling by FGFR2
100	RNA polymerase II transcribes snRNA genes
101	Constitutive Signaling by Aberrant PI3K in Cancer
102	Signaling by Insulin receptor
103	Signaling by MET
104	Signaling by NTRK1 (TRKA)
105	PI Metabolism
106	Transport of Mature Transcript to Cytoplasm

106	Transport of Mature Transcript to Cytoplasm
107	Fcgamma receptor (FCGR) dependent phagocytosis
108	TP53 Regulates Metabolic Genes
109	Signaling by FGFR
110	Regulation of mRNA stability by proteins that bind AU-rich elements
111	MAPK6/MAPK4 signaling
112	Amplification of signal from the kinetochores
113	Amplification of signal from unattached kinetochores via a MAD2 inhibitory signal
114	Downstream TCR signaling
115	Anchoring of the basal body to the plasma membrane
116	Processing of DNA double-strand break ends
117	VEGFA-VEGFR2 Pathway
118	Respiratory electron transport
119	Signaling by NTRKs
120	PI3K/AKT Signaling in Cancer
121	Signaling by TGF-beta family members
122	PI5P, PP2A and IER3 Regulate PI3K/AKT Signaling
123	Signaling by VEGF
124	Interleukin-4 and Interleukin-13 signaling
125	Negative regulation of the PI3K/AKT network
126	Mitotic Spindle Checkpoint
127	Signaling by the B Cell Receptor (BCR)
128	Selenoamino acid metabolism
129	Respiratory electron transport, ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins.
130	Metabolism of water-soluble vitamins and cofactors
131	Resolution of Sister Chromatid Cohesion
132	Cell-Cell communication
133	Host Interactions of HIV factors
134	HDR through Homologous Recombination (HRR) or Single Strand Annealing (SSA)
135	Fc epsilon receptor (FCERI) signaling
136	Cell surface interactions at the vascular wall
137	Homology Directed Repair
138	Late Phase of HIV Life Cycle
139	RHO GTPases Activate Formins
140	Influenza Life Cycle
141	HIV Life Cycle

142	Influenza Infection
143	DNA Double-Strand Break Repair
144	G2/M Checkpoints
145	The citric acid (TCA) cycle and respiratory electron transport
146	Apoptosis
147	Programmed Cell Death
148	mRNA Splicing - Major Pathway
149	Major pathway of rRNA processing in the nucleolus and cytosol
150	Metabolism of vitamins and cofactors
151	Separation of Sister Chromatids
152	mRNA Splicing
153	rRNA processing in the nucleus and cytosol
154	G2/M Transition
155	Mitotic G2-G2/M phases
156	Mitotic Prometaphase
157	Cilium Assembly
158	Mitotic Anaphase
159	Mitotic Metaphase and Anaphase
160	rRNA processing
161	Phospholipid metabolism
162	Ub-specific processing proteases

R Script

```
#PRE-PROCESSING


library(GEOquery); library(affy); library(limma); library(oligo); library(readr); library(ggplot2); library(ggpubr); library( ReactomePA); library(formattable); library(reactome.db);
library(huex10sttranscriptcluster.db)

#---------------------------
#uncomment to unpack CEL files
#set directory to file location

#untar("GSE143150_RAW.tar",list=TRUE)  ## check contents
#untar("GSE143150_RAW.tar")
#list.files(pattern="*.CEL.gz")
#file.rename(list.files(pattern="*.CEL.gz"), paste0("GSM4250",986:997,".CEL.gz"))
#list.files()
#---------------------------
#-------- Data extraction from files
#-------- Functions for annotating IDs and Graph plotting
#---------------------------

Annot <- data.frame(SYMBOL=sapply(contents(huex10sttranscriptclusterSYMBOL), paste, collapse=","),
                    GENENAME=sapply(contents(huex10sttranscriptclusterGENENAME), paste, collapse=","),
                    ENSEMBLID=sapply(contents(huex10sttranscriptclusterENSEMBL), paste, collapse=","),
                    ENTREZID=sapply(contents(huex10sttranscriptclusterENTREZID), paste,collapse=","))

gse <- getGEO('GSE143150',GSEMatrix = F)
genotype <- function(gsm) {
  Meta(gsm)[['characteristics_ch1']][2] #culture treatment is 3
}
treatment <- function(gsm) {
  Meta(gsm)[['characteristics_ch1']][3] #EMT promoted
}

annotate_id <- function(x) {
 y <-rownames(x)
 z <-as.character(as.factor(x$logFC))
 item <- matrix(data=NA, nrow=length(y), ncol=5)
 i <- 1
 for (ids in y) {
  item[i,1] <- ids
  temp <- Annot[grep(ids,rownames(Annot)),]
  item[i,2] <- as.character(temp[[1]])
  item[i,3] <- as.character(temp[[2]])
  item[i,4] <- as.character(temp[[3]])
  item[i,4] <- as.character(temp[[3]])
  p <- as.character(temp[[4]])
  item[i,5] <- p
  i <- i+1
 }
 colnames(item) <- c('transcript_cluster_id', 'SYMBOL','GENENAME','ENSEMBLID','ENTREZID')
 item <- as.data.frame(item)
 return(item)
}

findplot <- function(goi) {  #insert gene of interest to function in string
 goi <- goi
 df <-data.frame(Treatment=as.factor(sapply(GSMList(gse),treatment)),
                 Genotype=as.factor(sapply(GSMList(gse),genotype)),
                 Expression_Level = as.factor(eset[goi,])) #put GOI inside []
 df$Genotype<- as.factor(df$Genotype)
 levels(df$Genotype) <- c("KD","WT")
 df$Treatment<- as.factor(df$Treatment)
 levels(df$Treatment) <- c("DIFF","FULL")

 wtdiff <- mean(as.numeric(as.character(subset(df$Expression_Level, df$Treatment == 'DIFF' & df$Genotype == 'WT'))))
 wtfull <- mean(as.numeric(as.character(subset(df$Expression_Level, df$Treatment == 'FULL' & df$Genotype == 'WT'))))
 mdiff <- mean(as.numeric(as.character(subset(df$Expression_Level, df$Treatment == 'DIFF' & df$Genotype == 'KD'))))
 mfull <- mean(as.numeric(as.character(subset(df$Expression_Level, df$Treatment == 'FULL' & df$Genotype == 'KD'))))

 shell <-data.frame(Treatment=as.factor(c('DIFF','FULL','DIFF','FULL')),
                    Genotype=as.factor(c('WT','WT','KD','KD')),
                    Expression_Level = as.factor(c(wtdiff,wtfull,mdiff,mfull)))
 shell$Expression_Level <- as.numeric(as.character(shell$Expression_Level))
 shell$Expression_Level <- round(shell$Expression_Level ,digit=2)
 shell$Expression_Level<- as.factor(shell$Expression_Level)

 g <- ggplot(data=shell,
             aes(x=Genotype,y=Expression_Level,group=Treatment)) +
  geom_line(aes(color=Treatment)) +
  geom_point(aes(color=Treatment)) +
  ggtitle(paste(as.character(Annot[grep(goi,rownames(Annot)),][[2]])))
 g
}
#---------------------------
#---------------------------- Processing expression Data and applying eBayes
#---------------------------

apd <- data.frame(treatment=as.factor(sapply(GSMList(gse),treatment)),genotype=as.factor(sapply(GSMList(gse),genotype)))
apd$cond <- as.factor(paste(apd$treatment,apd$genotype,sep="_"))
levels(apd$cond) <- c("DIFF_KD","DIFF_WT","FULL_KD","FULL_WT")
acelfiles <- paste0(rownames(apd),'.CEL.gz')
data <- read.celfiles(acelfiles,phenoData = new("AnnotatedDataFrame",as.data.frame(apd)))


expression_data <- oligo::rma(data) # Background correction, Normalistation using rma() on dataset
eset <- exprs(expression_data)
model <- model.matrix( ~ 0 + expression_data$cond) #linear model, with intercept and the coefficient for all conditions ("DIFF_KD","DIFF_WT","FULL_KD","FULL_WT")
colnames(model) <- levels(expression_data$cond)
contrasts <- makeContrasts(DIFF_KD - DIFF_WT, #Contrast between genotypes(KD and WT) in DIFF medium
                           FULL_KD - FULL_WT, #Contrast between genotypes(KD and WT) in FULL medium
                           FULL_KD - DIFF_KD, #Contrast between media(DIFF and FULL) in KD genotype
                           FULL_WT - DIFF_WT, #Contrast between media(DIFF and FULL) in WT genotype
                           interaction=(DIFF_KD-DIFF_WT) - (FULL_KD - FULL_WT), #Contrast between different genotype with different media, also known as interaction
```

```
                (DIFF_KD - DIFF_WT) + ( FULL_KD - FULL_WT), #Contrast between genotypes across media
                (FULL_KD - DIFF_KD) + ( FULL_WT - DIFF_WT), #Contrast between media across genotypes
                levels = model)

expdata_fitted_contrasts <- lmFit(expression_data,model) #Expression data undergoes Empirical Bayes method w.r.t linear model
fitted.contrasts <- contrasts.fit(expdata_fitted_contrasts,contrasts) #Subsequently fitted with the 7 differnt contrasts above
fitted.aebayes <- eBayes(fitted.contrasts) #Dataset with Empirical Bayes method applied

true_gendiff <- topTable(fitted.aebayes,coef = 1,number=Inf,p.value = 0.05,lfc=1)
true_genfull <- topTable(fitted.aebayes,coef = 2,number=Inf,p.value = 0.05,lfc=1)
true_mediakd <- topTable(fitted.aebayes,coef = 3,number=Inf,p.value = 0.05,lfc=1)
true_mediawt <- topTable(fitted.aebayes,coef = 4,number=Inf,p.value = 0.05,lfc=1)
true_intxn <- topTable(fitted.aebayes,coef = 5,number=Inf,p.value = 0.05,lfc=1)
gen_in_allmedia <- topTable(fitted.aebayes,coef = 6,number=Inf,p.value = 0.05,lfc=1)
media_in_allgen <- topTable(fitted.aebayes,coef = 7,number=Inf,p.value = 0.05,lfc=1)


#---------------------------- #table 2
GOIS <- data.frame(Genotype_in_DIFF=nrow(true_gendiff),
                   Genotype_in_FULL=nrow(true_genfull),
                   Media_in_KD=nrow(true_mediakd),
                   Media_in_WT=nrow(true_mediawt),
                   Assuming_Interaction=nrow(true_intxn),
                   Genotypes_across_media=nrow(gen_in_allmedia),
                   Media_across_genotypes=nrow(media_in_allgen))
row.names(GOIS) <- c("Transcript clusters with p-values<0.05 and logFC>1")

#--- Mapped clusters that pass cutoff to their gene names and gene IDs using "huex10sttranscriptcluster.db", annotation file for "Affymetrix Human Exon 1.0 ST Array"
#------------ table 4a and 4b
# identify the differentially expressed genes found in KD across media
# identify the differentially expressed genes found in WT across media
#------------

media_in_allgen1= as.data.frame(annotate_id(media_in_allgen)) #MC1
true_mediakd1= as.data.frame(annotate_id(true_mediakd)) #MC2
true_mediawt1= as.data.frame(annotate_id(true_mediawt)) #MC3

unique_genes_kd <-subset(true_mediakd1, !(GENENAME %in% media_in_allgen1$GENENAME)) #Genes_KD
unique_genes_wt <-subset(true_mediawt1, !(GENENAME %in% media_in_allgen1$GENENAME)) #Genes_WT

#--------- Parsing Reactome
# find the pathways affected during KD

KD_after_R <- enrichPathway(unique_genes_kd$ENTREZID,organism = "human", pvalueCutoff = 1, readable = T)
gene_media_kd = summary(KD_after_R)
gene_media_kd <- subset(gene_media_kd, select=c('geneID','Description','GeneRatio','BgRatio','pvalue','Count')) #5
unique_pathways_kd<-as.data.frame(gene_media_kd$Description)
colnames(unique_pathways_kd) <- "pathways affected in KD cells"

#----------
# plotting of line plots for genes in X
#----------

cluster_X <- ggarrange(findplot("2555490"),findplot("2639054"),findplot("2813060"),
                       ncol = 3, nrow = 1)


#------------------------------------------ POST-PROCESSING
#------------------------------------------ POST-PROCESSING
#------------------------------------------ POST-PROCESSING

# number of transcripts meeting cut off for each contrast (2)
# Vplots  (3a,3b,3c) --> looking into 3 contrasts (1:Varying media in KD, 2: Varying media in WT, 3: Varying media across genotype)
# Mapped clusters that pass cutoff to their gene names and gene IDs using "huex10sttranscriptcluster.db", annotation file for "Affymetrix Human Exon 1.0 ST Array"
# identify the differentially expressed genes found in KD across media (4a) <20>
# identify the differentially expressed genes found in WT across media (4b) <2>
# Parsed REACTOME: Pathways affected in KD across media , WT across media, WT and KD across media (5a,5b,5c)
# Line plots of classified genes X (6)

formattable(GOIS) #table 2

#194 GOIs <MEDIA_KD> #table 3a
volcanoplot(fitted.aebayes,coef = 3, main=sprintf("%d features (Between media in KD) pass cutoff [LOG FOLD CHANGE >1 , P-VALUE<0.05]",nrow(true_mediakd))); points(true_mediakd[['logFC']],-
log10(true_mediakd[['P.Value']]),col='red')

#89 GOIs <MEDIA_WT> #table 3b
volcanoplot(fitted.aebayes,coef = 4, main=sprintf("%d features (Between media in WT) pass cutoff [LOG FOLD CHANGE >1 , P-VALUE<0.05]",nrow(true_mediawt))); points(true_mediawt[['logFC']],-
log10(true_mediawt[['P.Value']]),col='red')

#576 GOIs  #table 3c
volcanoplot(fitted.aebayes,coef = 7, main=sprintf("%d features (Between media across genotypes) pass cutoff [LOG FOLD CHANGE >1 , P-VALUE<0.05]",nrow(media_in_allgen)));
points(media_in_allgen[['logFC']],-log10(media_in_allgen[['P.Value']]),col='red')

formattable(unique_genes_kd)  #table 4a <20 genes>
formattable(unique_genes_wt) #table 4b <2 genes>

formattable(head(unique_pathways_kd,n=10)) #table 5 <162 pathways>
write.csv(unique_pathways_kd,"fig5.csv")

cluster_X #table 6
```