Data in Brief 48 (2023) 109069

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Transcriptomic data of bevacizumab-adapted colorectal adenocarcinoma cells HCT-116

Sala Cesare[a,1], Lottini Tiziano[a,1], Lastraioli Elena[a,b], Ruffinatti Federico Alessandro[c], Visentin Luca[c], Arcangeli Annarosa[a,b,*]

[a] Department of Experimental and Clinical Medicine, University of Florence, Viale GB Morgagni 50, 50134 Florence, Italy
[b] Complex Dynamics Study Centre (CSDC), University of Florence, 50100 Florence, Italy
[c] Department of Life Sciences and Systems Biology, University of Torino, Via Accademia Albertina 13, 10123 Torino, Italy

ARTICLE INFO

Article history:
Received 2 January 2023
Revised 28 February 2023
Accepted 10 March 2023
Available online 17 March 2023

Dataset link: Transcriptomic data of Bevacizumab-adapted colorectal adenocarcinoma cells HCT-116 (Original data)

Keywords:
Gene expression
HCT116
Bevacizumab
Bioinformatic analysis

ABSTRACT

A bioinformatic approach was applied to evaluate the effect of treatment with Bevacizumab on the gene expression profile of colorectal adenocarcinoma cells. The transcriptomic profile of Bevacizumab-adapted HCT-116 (Bev/A) colorectal adenocarcinoma cells was determined and compared with that of the corresponding control cell line by Agilent microarray analysis. Raw data were preprocessed, normalized, filtered, and subjected to a differential expression analysis using standard R/Bioconductor packages (i.e., limma, RankProd). As consequence of Bevacizumab adaptation, 166 differentially expressed genes (DEGs) emerged, most of them (123) resulted downregulated and 43 overexpressed. The list of statistically significant dysregulated genes was used as an input for functional overrepresentation analysis using ToppFun web tool. Such analysis pointed at cell adhesion, cell migration, extracellular matrix organization and angiogenesis as the main dysregulated biological process involved in Bevacizumab-adaptation of HCT116 cells. In addition, gene set enrichment analysis was performed using GSEA, searching for enriched terms within the Hallmarks (H), Canonical

* Corresponding author.
E-mail address: annarosa.arcangeli@unifi.it (A. Annarosa).
[1] Cesare Sala and Tiziano Lottini equally contributed to this work.

https://doi.org/10.1016/j.dib.2023.109069

Pathways (CP), and Gene Ontology (GO) gene sets. GO terms that showed significant enrichment included: transportome, vascularization, cell adhesion and cytoskeleton, extra cellular matrix (ECM), differentiation and epithelial–mesenchymal transition (EMT), inflammation and immune response. Raw and normalized microarray data were deposited in the Gene Expression Omnibus (GEO) public repository with accession number GSE221948.

Specifications Table

Subject	*Biology*
Specific subject area	Transcriptomics
Type of data	Tables Figures
How the data were acquired	Total RNA was extracted from the samples and checked for quality and integrity through Agilent 2100 Bioanalyzer, with RNA 6000 Nano kit (Agilent Technologies). Agilent-026652 Whole Human Genome Microarray 4 × 44K v2 platform (Agilent Technologies) was used to measure gene expression, following manufacturer's protocols. Agilent G49000 DA SureScan Microarray scanner and Agilent Feature Extraction software were used for image acquisition and data extraction, respectively. Raw data were processed in R/Bioconductor environment by means of a standard pipeline of analysis for microarray gene expression data.
Data format	Raw Analyzed Filtered
Description of data collection	We used a colorectal cancer cell line (HCT-116) as control and a colorectal cancer cell line adapted to bevacizumab (HCT-116 Bev/A) as biological model. The sample size was: HCT-116 $n = 5$; HCT-116 Bev/A $n = 5$. Gene expression was assessed using Agilent microarray technology. Raw data were background-subtracted, $\log_2$ transformed, and *quantile-quantile* normalized. Based on PCA results, a batch-effect correction was applied and one control sample (Ctrl_5) was removed from the dataset as an outlier.
Data source location	Institution: Department of Experimental and Clinical medicine, University of Florence City: Florence Country: Italy Latitude and longitude for collected samples/data: 43.9027681, 11.2463084
Data accessibility	Repository name: Gene Expression Omnibus (GEO) database Data accession number: GSE221948 Direct URL to data: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE221948

Value of the Data

- These data describing the transcriptomic profile of HCT-116 Bev/A can be used to identify deregulated genes related to Bevacizumab adaptation.
- Researchers working on drug resistance in colorectal carcinoma might find these data useful for identification of altered genes that might be targeted to overcome Bevacizumab resistance

- These genes might represent new targets for therapies aimed at overcoming the Bevacizumab resistance in CRCs. A better understanding in the intrinsic and acquired therapy resistance will be a great asset for drug development.

1. Objective

Due to the availability of various chemotherapy regimens, the overall survival of patients with advanced colon cancer has been improved over the past decades, particularly with the advent of monoclonal antibodies such as Bevacizumab. However, Chemo- and targeted therapies provide only a limited increase of overall survival for these patients, with drug resistance reportedly developing in nearly all patients with colon cancer, leading to a decrease in the therapeutic efficacies of anticancer agents, and finally to chemotherapy failure.

In order to characterize Bevacizumab-adapted colorectal cancer cells (HCT-116), we compared their transcriptomic profile with that of HCT116 control cells.

2. Data Description

Transcriptomic analysis was used to determine the expression profile of Bevacizumab-adapted HCT-116 (Bev/A) colorectal adenocarcinoma cells and compare it with that of the control, not adapted, cell line. The raw data were processed using R/Bioconductor programs for normalization, filtering, and differential expression analysis (i.e., limma, RankProd). Hierarchical clustering and Principal Component Analysis (PCA) were used as inter-sample distance metrics and an outlier sample (ctrl 5) was identified (Fig. 1). We next performed differential expression analysis (DEA) to identify differentially expressed genes (Fig. 2). Adaptation of HCT-116 cells to Bevacizumab resulted in 166 differentially expressed genes (DE), most of them (123) being downregulated and 43 up regulated (Table SM1). The top 20 upregulated and downregulated genes are reported in Table 1. HCT-116 cells adapted to bevacizumab showed the de-regulation of specific biological process such as cell adhesion, cell migration, extracellular matrix organization, and angiogenesis (Fig. 3 and Table SM2). Furthermore, a gene set enrichment analysis was conducted using GSEA software to search for enriched terms within Hallmarks (H), Canonical Pathways (CP), and Gene Ontology (GO) gene sets. The GO term analysis revealed several main functional categories which characterized adapted vs not adpted HCT-116 cells. These categories included transportome, vascularization, cell adhesion and cytoskeleton, extra cellular matrix (ECM), differentiation and epithelial-mesenchymal transition (EMT), inflammation, and immune response. (Fig. 4 and Table SM3).

3. Experimental Design, Materials and Methods

Cell lines. HCT-116 colorectal cancer cells (Ctrl group) and HCT-116-Bev/A (adapted to Bevacizumab, Res group) were kindly gifted by professor L.M. Ellis from the MD Anderson Cancer Center [1]. HCT-116 cells were cultured in RPMI (Invitrogen) added with 10% FBS (EuroClone) and maintained at 37°C and 5% CO_2 as in Lastraioli E. et al [2]. Cells were thawed two weeks before the RNA extraction and maintained at passage number 5 before RNA extraction. HCT-116-Bev/A were maintained adding Bevacizumab 250 µg/ml to the culture medium, according to Fan et al. [1]. Mycoplasma contamination was excluded in both cell lines by end point PCR, using Platinum PCR Supermix (Invitrogen, Life Technologies) and the following primers: mico down: 5'-TGCACCATCTGTCACTCTGTTAACCTC-3'; mico up: 5'-ACTCCTACGGGAGGCAGCAGTA-3'. PCR conditions were the following: denaturation at 94°C for 5 min, 35 cycles at 94°C for 1 min, 58°C for 1,30 min, 72°C for 1.30 min and a final extension cycle at 72°C for 10 min.

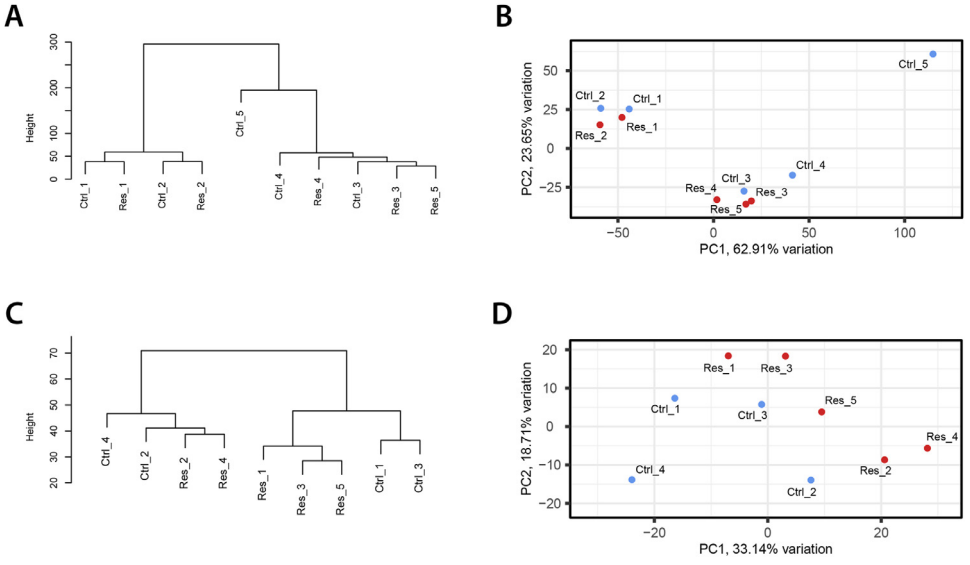

Fig. 1. Preliminary plots for sample quality control. Hierarchical clustering (Ward's minimum variance method with Euclidean distance measure) and Principal Component Analysis (PCA) were used as inter-sample distance metrics to detect possible outliers and batch effects (**A,B**) Dendrogram and PCA representing the full expression dataset in its initial form. Both representations clearly indicated the presence of an outlier sample (Ctrl_5) as well as a marked batch effect separating samples 1 and 2 from the others. (**C,D**) Dendrogram and PCA representing the expression dataset after outlier removal and batch effect correction. Overall, the procedure was quite effective (as shown by the much shorter height range of C compared to A) and enabled a good separation of the two experimental groups in the principal component space. Blue dots are the *Ctrl* group representing HCT-116 cells; Red dots are the *Res* group representing HCT-116-Bev/A (adapted to Bevacizumab) cells.

RNA extraction and quality control. Total RNA was extracted from the samples with Trizol reagent (Invitrogen) according to the manufacturer's protocol. The extracted RNA was then checked for its quality and integrity by Agilent 2100 Bioanalyzer with RNA 6000 Nano kit (Agilent Technologies), while RNA concentration was measured by Nanodrop ND-1000 (Thermo Scientific).

Microarray hybridization. Gene expression was assessed through one-channel microarray technology. RNA samples whose RIN was $>$ 6 were hybridized onto Agilent-026652 Whole Human Genome Microarray 4 $\times$ 44K v2 (Agilent Technologies) according to the manufacturer's protocol. Fluorescence signal was then acquired through an Agilent G49000 DA SureScan Microarray scanner and row data were obtained using the Agilent Feature Extraction software.

Differential expression analysis. Raw data were analyzed running a custom script in R/Bioconductor environment implementing a standard pipeline for gene expression data analysis that featured the following steps:

1. *Norm-exp* method was used to subtract the background signal.
2. Data were $\log_2$-transformed.
3. *Quantile–quantile* procedure was used for inter-array-normalization.
4. Hierarchical clustering (Ward's minimum variance method with Euclidean distance measure) and Principal Component Analysis (PCA) of the samples were used for quality control and batch effect detection (Fig. 1A, B).
5. Based on these inter-sample metrics, one sample from the control group (Ctrl_5) was removed as an outlier and the *removeBatchEffect()* function from the *limma* package (v3.48.3) was used to mitigate the batch effect (Fig. 1C,D).

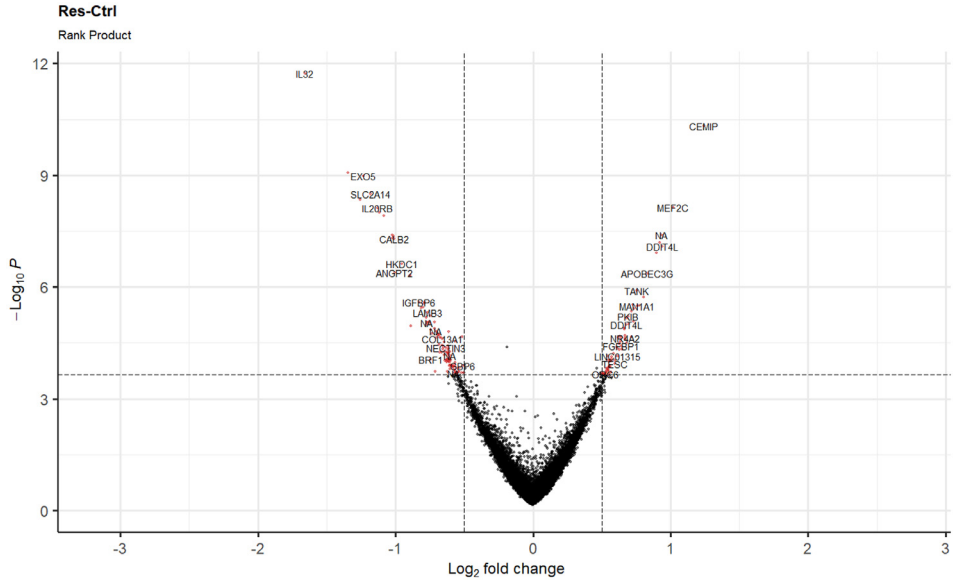

Fig. 2. Volcano plot showing the results of the differential expression analysis (DEA). DEA was carried out using the RankProd Bioconductor R package to perform a rank product-based statistical analysis. For each unfiltered probe in the array, the $-\log_{10}(p\text{-value})$ was plotted against its $\log_2$FC in *Res* as compared to *Ctrl* group. Statistically significant DEGs (i.e., genes with an adjusted p-value < 0.05 and $|\log_2 FC| > 0.5$) are shown in red. Genes with a positive $\log_2$FC were upregulated in HCT-116-Bev/A compared to HCT-116 control cells.

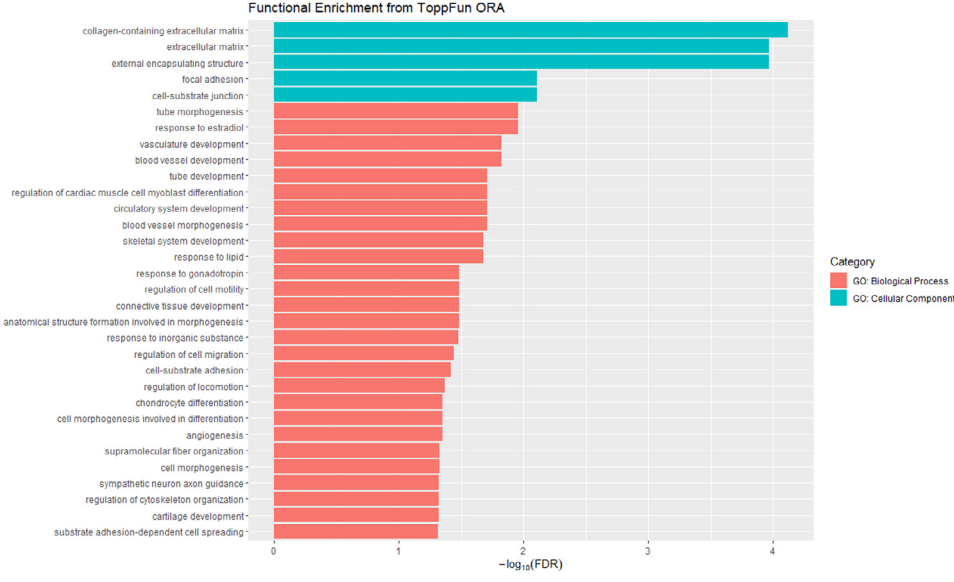

Fig. 3. Bar chart showing the statistically significant functional terms as returned by the OverRepresentation Analysis (ORA) of our DEG lists (ToppFun by ToppGene Suite, https://toppgene.cchmc.org/, accessed on 25 December 2022). Statistically significant (q-value < 0.05) functional terms related to the Gene Ontology (GO) of Cellular Components (blue bars) and Biological Processes (red bars) have been ranked according to their q-value, represented in the graph as bar length in terms of $-\log_{10}$(FDR). Overall, cell adhesion, cell migration, extracellular matrix organization, and angiogenesis were the main dysregulated biological processes. Details about the precise composition of the enriched gene sets can be found in Table SM2.

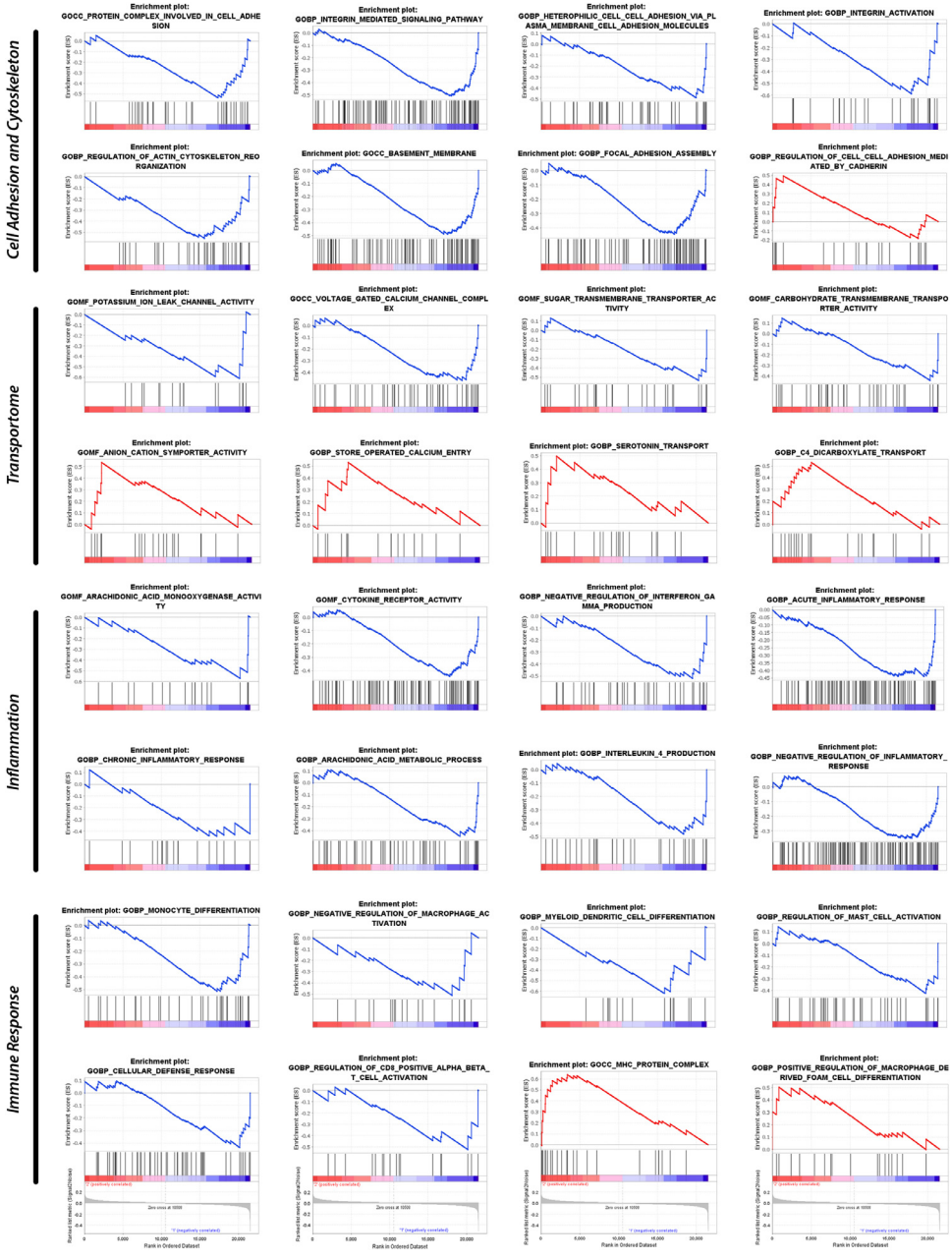

Fig. 4. Results of Gene Set Enrichment Analysis (GSEA software v4.2.2 and MSigDB database v7.4) are shown here as a selection of enrichment plots for some of the most representative GO terms belonging to four different functional categories: cell adhesion and cytoskeleton, transportome, inflammation, and immune response. Normalized Enriched Score (NES) for down-regulated terms (the majority) are in blue, while up-regulated terms are in red. In any case, NES was computed using the standard (weighted) scoring method. For a more complete list of all statistically significant (q-value < 0.25) enriched terms returned by GSEA see Table SM3.

Table 1
List of the top 20 up-regulated and down-regulated genes obtained by the rank product-based statistical analysis (adjusted p-value < 0.05 and $|\log_2 FC| > 0.5$). For each significant probe/gene the following information is reported: Agilent probe ID, gene symbol, $\log_2$ fold change, nominal (unadjusted) p-value, FDR q-value (or the percentage of false positives, pfp).

Agilent probe id	Gene symbol	Log2FC	P value	Adjusted P value
Top 20 up-regulated genes				
A_23_P324754	CEMIP	1,253257284	2,713E-11	7,937E-07
A_23_P320739	MEF2C	1,02385636	8,604E-09	0,0001258
A_23_P302672	DDIT4L	0,946193556	3,915E-08	0,0002863
A_33_P3220911	NA	0,944526664	3,005E-08	0,000293
A_23_P209978	VSNL1	0,91975167	8,469E-08	0,0004129
A_23_P58328	ANXA10	0,917297411	6,306E-08	0,000369
A_23_P143713	APOBEC3G	0,84295629	2,758E-07	0,001153
A_23_P39465	BST2	0,814767746	7,572E-07	0,002769
A_23_P39955	ACTG2	0,786372761	0,000001154	0,003752
A_33_P3303414	MAN1A1	0,759014096	0,000001759	0,005146
A_23_P154306	TANK	0,75389599	0,000001955	0,005199
A_33_P3407880	ANKRD22	0,752193982	0,000002572	0,005016
A_33_P3238166	PXDN	0,728096294	0,000002148	0,005238
A_33_P3271455	PXDN	0,725469955	0,000002379	0,005353
A_23_P209564	CYBRD1	0,719281186	0,000004012	0,007336
A_23_P85441	IGSF9	0,714301874	0,000002396	0,005006
A_33_P3390057	TM4SF1	0,713355379	0,00000732	0,01071
A_23_P145529	PKIB	0,695722612	0,000005601	0,009639
A_32_P40288	TMEM200A	0,682695932	0,000007291	0,01123
A_33_P3372099	DDIT4L	0,675765438	0,00000677	0,011
Top 20 down-regulated genes				
A_23_P15146	IL32	-1,687060688	1,077E-12	3,15E-08
A_23_P156327	TGFBI	-1,370722275	1,544E-10	0,000002259
A_23_P161190	VIM	-1,259423152	4,247E-09	0,00002071
A_23_P193	EXO5	-1,250961574	1,021E-09	0,000009954
A_32_P47754	SLC2A14	-1,189033824	1,46E-09	0,00001068
A_23_P91850	IL20RB	-1,138159145	3,603E-09	0,00002108
A_23_P1473	PRF1	-1,12763328	6,488E-09	0,00002712
A_23_P369994	DCLK1	-1,092883966	7,356E-09	0,0000269
A_23_P408285	PRICKLE1	-1,029982866	2,302E-08	0,00007482
A_23_P4714	MIA	-1,025028794	6,434E-08	0,0001448
A_23_P21092	CALB2	-1,020768865	4,542E-08	0,0001107
A_23_P60079	ANGPT2	-1,01649632	3,198E-08	0,00008506
A_33_P3268304	LIMS2	-1,013640911	2,313E-08	0,00006767
A_23_P202427	HKDC1	-0,970853654	1,137E-07	0,0002376
A_33_P3358208	PADI1	-0,916858765	2,993E-07	0,0005473
A_33_P3346826	IL32	-0,913033	0,00000026	0,000507
A_33_P3309911	PRAMEF5	-0,896853073	0,00001907	0,01268
A_23_P139912	IGFBP6	-0,826192536	0,000001228	0,001996
A_23_P3552	RRN3P1	-0,818850561	0,00000568	0,005934
A_33_P3211804	RUNX1	-0,801572569	0,00000231	0,003379

6. Expression values were filtered to discard unexpressed genes. Only probes featuring a $\log_2$ expression above 6.5 in at least 75% of the samples of at least one of the two groups were retained. Overall, 29,254 probes out of 34,127 (85.72%) were kept for subsequent analysis.
7. Within-group mean expression values were used for $\log_2$ fold change (FC) computation.
8. Statistical significance was assigned to each probe based on the rank product statistic as returned by the *RP.advance()* function from the *RankProd* package (v3.18.0) [3–5].
9. *P*-values were adjusted for multiple comparisons using the Benjamini–Hochberg procedure to control the false discovery rate (BH-FDR) at level $\alpha = 0.05$ [6].
10. An additional cutoff on FCs was applied removing from the list of DEGs all those probes with $|\log_2 FC| < 0.5$.

11. The final DEG list (Table SM1) featured 166 significant probes (123 with a negative and 43 with a positive $\log_2$FC), representing 135 unique annotated genes, of which 97 were down- and 38 up-regulated (see Fig. 2 and Table 1).

Functional Enrichment analysis. ToppFun web tool (by ToppGene Suite, https://toppgene.cchmc.org/, accessed on 25 December 2022) was used for functional OverRepresentation Analysis (ORA) of the DEG list [7]. Based on the results of the hypergeometric hypothesis test, all terms with a BH-FDR q-value < 0.05 were considered statistically significant (see Fig. 3 and Table SM2). In addition, Gene Set Enrichment Analysis (GSEA) was performed using the GSEA software v4.2.2 with the MSigDB database v7.4 [8]. Microarray probes were collapsed into unique gene symbols before the analysis, and a standard (weighted) enrichment statistic was chosen for Normalized Enriched Score (NES) computation. Within the context of the GSEA, the threshold on the q-value for a gene set to be considered statistically significant was set to 0.25 (see Fig. 4 and Table SM3).

Ethics Statements

Not applicable.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

Transcriptomic data of Bevacizumab-adapted colorectal adenocarcinoma cells HCT-116 (Original data) (Gene Expression Omnibus (GEO)).

CRediT Author Statement

Sala Cesare: Investigation, Formal analysis, Methodology, Software, Data curation, Validation, Visualization, Writing – original draft; **Lottini Tiziano:** Investigation, Methodology, Data curation, Writing – original draft; **Lastraioli Elena:** Investigation, Methodology, Data curation, Visualization, Writing – original draft, Writing – review & editing; **Ruffinatti Federico Alessandro:** Investigation, Formal analysis, Methodology, Software, Data curation, Validation, Writing – original draft, Writing – review & editing; **Visentin Luca:** Investigation, Formal analysis, Methodology, Software, Data curation, Validation; **Arcangeli Annarosa:** Conceptualization, Supervision, Project administration, Funding acquisition, Writing – original draft, Writing – review & editing.

Acknowledgments

This work was supported by Associazione Italiana per la Ricerca sul Cancro (AIRC, grant no. 1662, 15627 and IG 21510) to A.A.; PRIN Italian Ministry of University and Research (MIUR) “Leveraging Basic Knowledge of the Ion Channel Network in Cancer For Innovative Therapeutic Strategies (LIONESS)” 20174TB8KW to A.A. pHioniC European Union’s Horizon 2020 grant No 813834 to A.A.

Supplementary Materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.dib.2023.109069.

References

[1] F. Fan, S. Samuel, P. Gaur, J. Lu, N.A. Dallas, L. Xia, D. Bose, V. Ramachandran, L.M. Ellis, Chronic exposure of colorectal cancer cells to bevacizumab promotes compensatory pathways that mediate tumour cell migration, Br. J. Cancer 104 (8) (2011) 1270–1277, doi:10.1038/bjc.2011.81.
[2] E. Lastraioli, L. Guasti, O. Crociani, S. Polvani, G. Hofmann, H. Witchel, L. Bencini, M. Calistri, L. Messerini, M. Scatizzi, R. Moretti, E. Wanke, M. Olivotto, G. Mugnai, A. Arcangeli, herg1 gene and HERG1 protein are overexpressed in colorectal cancers and regulate cell invasion of tumor cells, Cancer Res. 64 (2) (2004) 606–611.
[3] R. Breitling, P. Armengaud, A. Amtmann, P. Herzyk, Rank products: A simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments, FEBS Lett. 573 (2004) 83–92.
[4] F. Hong, R. Breitling, C.W. McEntee, B.S. Wittner, J.L. Nemhauser, J. Chory, RankProd: A bioconductor package for detecting differentially expressed genes in meta-analysis, Bioinformatics 22 (2006) 2825–2827.
[5] F. Del Carratore, A. Jankevics, R. Eisinga, T. Heskes, F. Hong, R. Breitling, RankProd 2.0: A refactored bioconductor package for detecting differentially expressed features in molecular profiling datasets, Bioinformatics 33 (2017) 2774–2775.
[6] Y. Benjamini, D. Drai, G. Elmer, N. Kafkafi, I. Golani, Controlling the false discovery rate in behavior genetics research, Behav. Brain Res. 125 (2001) 279–284.
[7] J. Chen, E.E. Bardes, B.J. Aronow, A.G. Jegga, ToppGene Suite for gene list enrichment analysis and candidate gene prioritization, Nucleic. Acids. Res. 37 (2009) W305–W311.
[8] A. Subramanian, P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette, A. Paulovich, S.L. Pomeroy, T.R. Golub, E.S. Lander, et al., in: Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles, 102, 2005, pp. 15545–15550.