Data in Brief 34 (2021) 106643

Contents lists available at ScienceDirect

Data in Brief

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

Data Article

RNA sequencing data of different grade astrocytoma cell lines

Juliana Ferreira de Sousa[a], Patrick da Silva[b], Rodolfo Bortolozo Serafim[b,c], Ricardo Perecin Nociti[c,d], Cristiano Gallina Moreira[b], Wilson Araujo Silva[c,e,f], Valeria Valente[b,c,*]

[a] *Radiation Oncology Branch, National Cancer Institute, NIH, 10 Center Drive, Bethesda, MD 20892, United States*
[b] *Department of Clinical Analysis, School of Pharmaceutical Sciences, São Paulo State University (UNESP), Rodovia Araraquara-Jaú, Km 01 - s/n, Campos Ville, Araraquara, SP 14800-903, Brazil*
[c] *National Institute of Science and Technology in Stem Cell and Cell Therapy and Center for Cell-Based Therapy, Rua Tenente Catão Roxo, 2501, Monte Alegre, Ribeirão Preto, SP 14051-140, Brazil*
[d] *Department of Veterinary Medicine, Faculty of Animal Sciences and Food Engineering, University of São Paulo (USP), Av. Duque de Caxias Norte, 225, Campus Fernando Costa, 13635-900 Pirassununga, SP, Brazil*
[e] *Department of Genetics, Ribeirão Preto Medical School, University of São Paulo (USP), Avenida Bandeirantes, 3900, 14049-900 Ribeirão Preto, SP, Brazil*
[f] *Center for Integrative Systems Biology, CISBi, NAP/USP, Rua Catão Roxo, 2501, Monte Alegre, Ribeirão Preto, SP 14051-140, Brazil*

ARTICLE INFO

Article history:
Received 10 August 2020
Revised 7 December 2020
Accepted 8 December 2020
Available online 10 December 2020

Keywords:
RNAseq
Glioblastoma
Astrocytoma
Tumor progression
Gene expression profiling
Replicative stress
Camptothecin (CPT)

ABSTRACT

Astrocytomas are the most common and aggressive type of primary brain tumors in adults. The World Health Organization (WHO) assorts them into grades, from I to IV, based on histopathological features that reflect their malignancy [1]. Alongside with tumor progression, comes an increased proliferation, genomic instability, infiltration in normal brain tissue and resistance to treatments. The high genomic instability forges tumor cells enhancing key proteins that avoid cells from collapsing and favor therapy resistance [2]. To explore genes and pathways associated with tumor progression phenotypes we analyzed gene expression in a panel of non-tumor and glioma cell lines, namely: ACBRI371, non-tumor human astrocytes; HDPC, fibroblasts derived from

* Corresponding author at: Department of Clinical Analysis, School of Pharmaceutical Sciences, São Paulo State University (UNESP), Rodovia Araraquara-Jaú, Km 01 - s/n, Campos Ville, Araraquara, SP 14800-903, Brazil.
E-mail address: valeria.valente@unesp.br (V. Valente).
Social media: (J.F. de Sousa), (R.P. Nociti), (C.G. Moreira)

https://doi.org/10.1016/j.dib.2020.106643
2352-3409/© 2020 Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

dental pulp; Res186, Res259, Res286 and UW467 that include grade I, II and III astrocytoma cell lines derived from pediatric tumors; and T98G, U343MG, U87MG, U138MG and U251MG, all derived from GBM (grade IV). We also profiled gene expression changes caused by exogenously induced replicative stress, performing RNA sequencing with camptothecin (CPT)-treated cells. Here we describe the RNA-sequencing data set acquired, including quality of reads and sequencing consistency, as well as the bioinformatics strategy used to analyze it. We also compared gene expression patterns and pathway enrichment between non-tumor *versus* lower-grade (LGG), non-tumor *versus* GBM, LGG *versus* GBM, and CPT-treated *versus* non-treated cells. In brief, a total of 6467 genes showed differential expression and 5 pathways were enriched in tumor progression, while 2279 genes and 7 pathways were altered under the replication stress condition. The raw data was deposited in the NCBI BioProject database under the accession number PRJNA631805. Our dataset is valuable for researchers interested in differential gene expression among different astrocytoma grades and in expression changes caused by replicative stress, facilitating studies that seek novel biomarkers of glioma progression and treatment resistance.

© 2020 Published by Elsevier Inc.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

Specifications Table

Subject	Cancer Research
Specific subject area	Transcriptomic changes in different grade astrocytoma cells, comparing gene expression changes that occur in tumor progression or under replicative stress induced by the topoisomerase I inhibitor, Camptothecin (CPT).
Type of data	Transcriptomic data Figures Tables
How data were acquired	RNA sequencing: Bioanalyzer Instrument (Agilent), lllumina Sequencers Genome Analyzer IIx and NextSeq 500 (Illumina Inc.). Software: FastQC, Trimmomatic, SGA, HISAT2, SAMtools, HTSeq-count, DESeq2
Data format	Raw: sra file format (repository link below) Analyzed: excel spreadsheet, tif figures
Parameters for data collection	Cells were grown under standard conditions or treated with CPT for 18 h, then RNA isolation and sequencing were performed.
Description of data collection	Total RNA was isolated using RNeasy mini kit (Qiagen), RNA quality was evaluated by Bioanalyzer (Agilent), rRNA was removed from samples and then samples were clustered and sequenced.
Data source location	Institution: Ribeirão Preto Blood Bank Ribeirão Preto, São Paulo, Brazil Coordinates: 21°11′18.1″S 47°48′17.3″W (−21.188357, −47.804813)
Data accessibility	Raw data is available at NCBI BioProject repository under the identification number PRJNA631805. Direct URL to data: https://www.ncbi.nlm.nih.gov/bioproject/PRJNA631805

Value of the Data

- These data provide essential information on gene expression profiling of normal astrocytes and different astrocytoma cell lines, and of GBM cells submitted to CPT-induced replicative stress.
- Scientists who study gene expression regarding astrocytoma progression and its resistance to genotoxic treatments would benefit from this data set to expand their knowledge and apply new insights to the current clinical management of patients.
- The dataset generated here can be used to design new experiments and projects aiming to use the analyzed cell lines as a model to improve disease progression understanding.
- This RNA-sequencing dataset can be further explored for the identification of novel biomarkers of prognosis prediction and/or treatment responsiveness.
- This dataset can also be interrogated intending the identification of potential new target-genes for the development of new drugs and/or therapeutic approaches that sensitize tumor cells to the available treatments.

1. Data Description

In this report we present the RNA sequencing analysis of different grade astrocytoma cell lines. Astrocytomas are the most common and aggressive type of primary brain tumors in adults. The World Health Organization (WHO) assorts them into grades, from I to IV, based on histopathological features that reflect their malignancy [1]. Grade I comprise benign curable tumors that are more frequent in children. Grade II are considered low-grade lesions, with restrained mitotic activity, but showing infiltrative capacity and tendency to progress to higher grades. Grade III present prominent mitotic activity, nuclear atypia and are also prone to undergo progression to grade IV. Glioblastoma (GBM), the most aggressive type of astrocytoma, is classified as grade IV and exhibits considerably higher mitotic activity, atypia, likewise angiogenesis and necrosis [1,2]. To explore genes and pathways associated with tumor progression we analyzed gene expression in a panel of non-tumor and glioma cell lines. We used two non-tumor cells and 9 cell lines representative of tumor progression, comprising: two non-tumor cells (HDPC and ACBRI371), two grade I (Res186, Res286), one grade II (Res259), one grade III (UW467) astrocytoma cells, and 5 GBM (T98G, U343MG, U87MG, U138MG and U251MG) cell lines. Here we considered cell lines from grades I, II and III as a group representing lower-grade glioma (LGG) and the GBM cell lines as representative of higher-grade glioma (HGG). Additionally, we evaluated the impact of CPT-induced replication stress in the transcriptome of two GBM cells, the most resistant (U138MG) and the most sensitive (U251MG) (data not shown), along with non-tumor control cells. Data collection was obtained in two rounds of sequencing (with Genome Analyzer IIx and with NextSeq 500, Illumina Inc.), to increment the total amount of reads produced and complete the sample set to be studied. We generated from 27.2 to 33.6 million of reads (trimmed/aligned) for libraries sequenced in the first run and from 58.2 to 86.1 million of reads for libraries sequenced in the second run (Table 1). With this dataset we could measure the expression levels of 44,608 genes among all samples evaluated. To verify the consistency of the generated data, we made scatter plots with the number of reads obtained per gene whose expression was detected in each sequenced sample (Fig. 1). Plots depicted the maximum Pearson correlation coefficient (PCC) when comparing different datasets (labeled A, B, C or D) of the same sample, in both rounds of sequencing (#1 and #2), for all groups of cells analyzed: non-tumor (Fig. 1A), LGG (Fig. 1B), GBM (Fig. 1C) and CPT treated GBM cells (Fig. 1D). Among non-tumor cells, we observed decreased PCC values when comparing ACBRI371 cells treated or not with CPT (0.87–0.88) (Fig. 1A). HDPC datasets showed a significant lower correlation with ACBRI371 cells, with PCC varying from 0.45 to 0.51 for the different conditions evaluated (Fig. 1A). In contrast, we detected a high degree of similarity concerning all LGG cells that showed PCC values varying in between 0.98 and 0.99 amongst all comparisons (Fig. 1B).

Table 1
RNA Sequencing metrics. The total amounts of reads produced for each cell line or condition analyzed were grouped into two datasets (A and B) for the first run (#1) or four datasets (A, B, C and D) for the second run (#2). Counting of the number of raw reads, timmmed reads and aligned reads are shown.

First run #1					
Sample	Datasets*	raw reads	trimmed reads	aligned reads	total aligned reads per condition
ACBRI371.A	A	18,777,557	16,073,974	15,938,790	32,064,952
ACBRI371.B	B	18,812,311	16,251,936	16,126,162	
HDPC.A	A	15,943,724	13,606,444	13,504,133	27,185,197
HDPC.B	B	15,979,931	13,774,367	13,681,064	
T98.A	A	17,728,147	15,076,520	14,976,750	30,171,998
T98.B	B	17,785,963	15,286,117	15,195,248	
U138.A	A	16,704,458	14,166,959	14,071,552	28,340,144
U138.B	B	16,744,624	14,355,064	14,268,592	
U251.A	A	18,962,975	16,244,799	16,142,384	32,480,625
U251.B	B	19,015,489	16,431,033	16,338,241	
U343.A	A	19,758,340	16,810,865	16,692,122	33,601,621
U343.B	B	19,791,840	17,016,541	16,909,499	
U87.A	A	19,764,536	16,898,823	16,788,200	32,998,714
U87.B	B	18,881,408	16,305,409	16,210,514	
Second run #2					
ACBRI371 + cpt18hs.A	A	21,518,098	21,095,090	20,862,592	83,405,190
ACBRI371 + cpt18hs.B	B	21,195,306	20,755,940	20,502,694	
ACBRI371 + cpt18hs.C	C	21,913,608	21,476,156	21,245,612	
ACBRI371 + cpt18hs.D	D	21,516,074	21,073,316	20,794,292	
R186.A	A	19,595,942	19,245,970	18,417,842	73,482,118
R186.B	B	19,230,796	18,859,758	18,014,710	
R186.C	C	19,960,034	19,593,568	18,755,774	
R186.D	D	19,544,228	19,171,266	18,293,792	
R259.A	A	18,566,902	18,175,510	16,400,986	65,617,994
R259.B	B	18,311,806	17,885,020	16,128,140	
R259.C	C	18,900,300	18,492,500	16,693,170	
R259.D	D	18,628,414	18,200,300	16,395,698	
R286.A	A	19,534,344	19,152,002	17,501,632	69,847,434
R286.B	B	19,192,134	18,795,248	17,144,668	
R286.C	C	19,876,010	19,479,954	17,808,962	
R286.D	D	19,487,048	19,087,884	17,392,172	
U138 + cpt18hs.A	A	22,201,702	21,832,016	14,562,900	58,244,234
U138 + cpt18hs.B	B	21,898,446	21,490,134	14,317,296	
U138 + cpt18hs.C	C	22,596,730	22,211,560	14,821,870	
U138 + cpt18hs.D	D	22,234,620	21,823,154	14,542,168	
U251 + cpt18hs.A	A	22,296,224	21,818,672	21,557,698	86,098,692
U251 + cpt18hs.B	B	21,929,756	21,428,122	21,135,756	
U251 + cpt18hs.C	C	22,702,128	22,210,046	21,951,788	
U251 + cpt18hs.D	D	22,280,758	21,776,944	21,453,450	
UW467.A	A	20,931,108	20,520,954	18,115,334	72,368,544
UW467.B	B	20,578,786	20,146,762	17,756,036	
UW467.C	C	21,313,478	20,886,716	18,446,260	
UW467.D	D	20,930,310	20,494,222	18,050,914	

*Refers to groups of reads obtained from different lanes of sequencing runs.

Much larger variation was observed for the GBM cell lines, in which PCC values remained at 0.62 or 0.64 in all comparisons (Fig. 1C). When GBM cells were exposed to CPT, we also observed a reduction in gene expression correlation between treated and non-treated cells, similarly to ACBRI371 cells. However, differences were more pronounced for U138MG (PCC=0.79) than for U251MG (PCC=0.97) cells (Fig. 1D).

Differential gene expression detected in the comparisons between the datasets representative of astrocytes, LGG and GBM cell lines are illustrated by the Volcano plots in Fig. 2. We

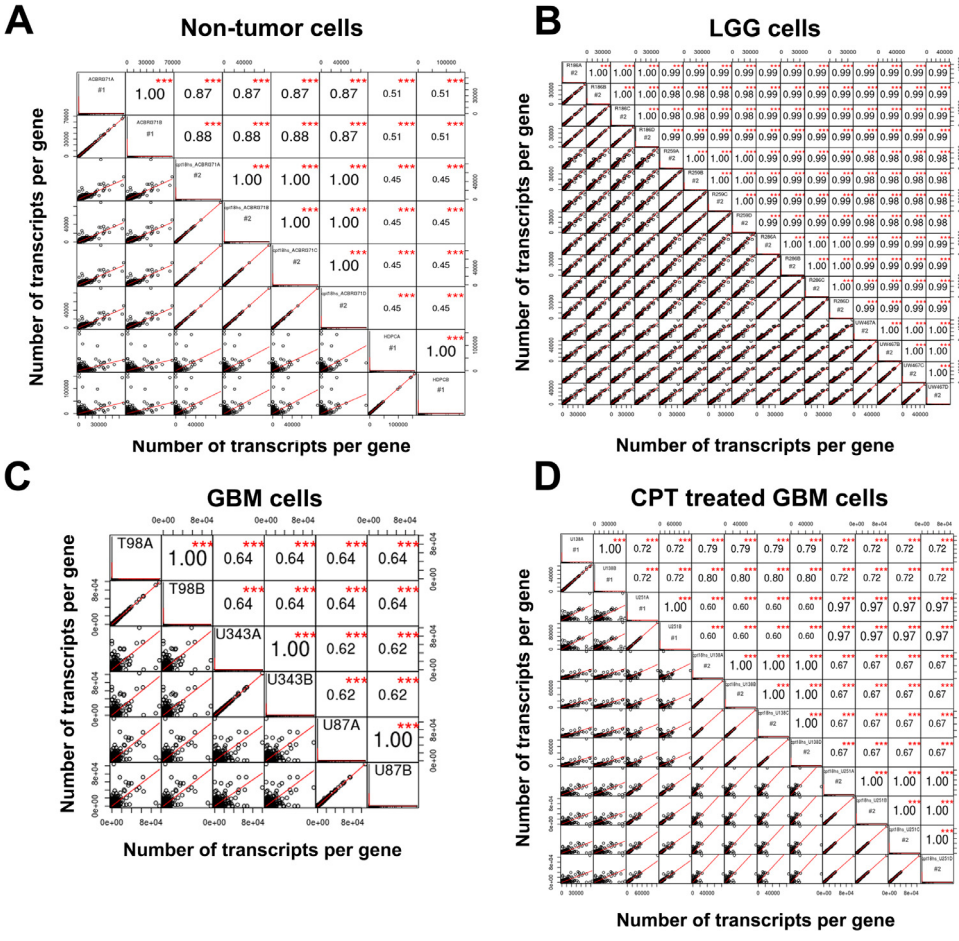

Fig. 1. Scatter plots showing the correlations among datasets of reads for each sample analyzed. Scatter plots were generated with the number of reads per gene for all genes whose expression was identified in each sequenced sample. Plots were clustered into four different groups, according to the cell line origin and/or treatment condition: non-tumor cells, LGG cells, GBM cells and CPT-treated GBM cells. #1 refers to samples sequenced in the Genome Analyzer IIx, while #2 refers to samples sequenced in the NextSeq 500, Illumina Inc. The letters A, B, C and D, accompanying the name of each cell line, refer to reads obtained from different lanes of sequencing for the same sample.

identified a total of 2877 genes differentially expressed between the groups of cells that characterize the progression from LGG to GBM, of which 1698 were down regulated and 1179 were up regulated (Fig. 2A). We also found 2466 altered genes by comparing ACBRI371 and LGG, being 1509 down regulated and 957 up regulated (Fig. 2B), and 1124 differentially regulated genes between ACBRI371 and GBM, being 250 up regulated and 874 down regulated (Fig. 2C). Among the cell lines submitted to a replicative stress condition, we identified 790 genes down regulated and 437 up regulated in ACBRI371 (Fig. 2D), 365 down regulated and 494 unregulated genes in U138MG (Fig. 2E) and 21 down regulated and 172 up regulated genes in U251MG (Fig. 2F). According to KEGG analysis, when considering all the altered genes found, we encountered enriched pathways only for: LGG *versus* GBM, ACBRI371 *versus* LGG, ACBRI371 *versus* GBM, and in ACBRI371 and U138MG cells CPT-treated *versus* non-treated (Table 2 and Fig. 3). For comparisons representative of tumor progression, the most preeminent pathways were pathways in cancer, PI3K-Akt signaling, neuroactive ligand-receptor interaction, cell adhesion molecules and calcium

Table 2
KEGG pathways enrichment analysis. Differentially expressed genes (q-values ≤ 0.0001 and log fold change > 2 or < −2) were submitted to KEGG evaluation. Enriched pathways found in each comparison and details of the analysis results are shown.

DOWN LGG - UP GBM								
Gene Set	Description	Size	Expect	Ratio	p value	FDR	LOGFOLD > 2	Enrichment%
hsa05200	Pathways in cancer	526	28.24	1.8059	0.00002066	0.001347	51	9.69581749
hsa04151	PI3K-Akt signaling pathway	354	19.006	2.1572	2.0102E-06	0.00021844	41	11.5819209
hsa05165	Human papillomavirus infection	339	18.2	1.978	0.000058394	0.0027195	36	10.61946903
hsa04510	Focal adhesion	199	10.684	2.8079	2.13E-07	0.000069599	30	15.07537688
hsa04010	MAPK signaling pathway	295	15.838	1.8942	0.00051407	0.016759	30	10.16949153
hsa04514	Cell adhesion molecules (CAMs)	144	7.7312	2.5869	0.000078204	0.0031868	20	13.88888889
hsa04512	ECM-receptor interaction	82	4.4025	3.8615	1.1754E-06	0.00019159	17	20.73170732
hsa04933	AGE-RAGE signaling pathway in diabetic complications	99	5.3152	3.1984	0.000017218	0.001347	17	17.17171717
hsa04668	TNF signaling pathway	110	5.9057	2.7092	0.00023505	0.0085141	16	14.54545455
hsa05144	Malaria	49	2.6307	4.1813	0.000042617	0.0023155	11	22.44897959
DOWN GBM - UP LGG	No enriched pathway							
DOWN ACBRI371 - UP GBM	No enriched pathway							
DOWN GBM - UP ACBRI371								
hsa04080	Neuroactive ligand-receptor interaction	277	10.866	2.1166	0.00051907	0.022	23	8.303249097
hsa04020	Calcium signaling pathway	183	7.1789	2.9253	8.7111E-06	0.0014199	21	11.47540984
hsa04723	Retrograde endocannabinoid signaling	148	5.8059	2.7558	0.00021105	0.01376	16	10.81081081
hsa04713	Circadian entrainment	96	3.766	3.7175	0.000019704	0.0021412	14	14.58333333
hsa05032	Morphine addiction	91	3.5698	3.6416	0.000048805	0.0039776	13	14.28571429
hsa04724	Glutamatergic synapse	114	4.4721	2.9069	0.00049148	0.022	13	11.40350877
hsa04727	GABAergic synapse	88	3.4521	3.1864	0.00060736	0.022	11	12.5
hsa05033	Nicotine addiction	40	1.5692	6.3729	2.2051E-06	0.00071887	10	25
hsa05133	Pertussis	76	2.9814	3.3541	0.00071129	0.023188	10	13.15789474
hsa05144	Malaria	49	1.9222	4.1619	0.00056511	0.022	8	16.32653061
DOWN ACBRI371 - UP LGG	No enriched pathway							

(*continued on next page*)

Table 2 (*continued*)

DOWN LGG - UP GBM								
Gene Set	Description	Size	Expect	Ratio	p value	FDR	LOGFOLD > 2	Enrichment%
DOWN LGG - UP ACBRI371								
hsa04080	Neuroactive ligand-receptor interaction	277	20.88	1.8678	0.000094512	0.0034882	39	14.07942238
hsa04514	Cell adhesion molecules (CAMs)	144	10.854	3.3166	6.42E-11	2.09E-08	36	25
hsa04510	Focal adhesion	199	15	2.0666	0.000076063	0.0034882	31	15.57788945
hsa04512	ECM-receptor interaction	82	6.181	3.3975	4.26E-07	0.000065465	21	25.6097561
hsa05032	Morphine addiction	91	6.8594	3.0615	2.6922E-06	0.00021942	21	23.07692308
hsa04012	ErbB signaling pathway	85	6.4071	2.9654	0.000013146	0.00085709	19	22.35294118
hsa05133	Pertussis	76	5.7287	2.9675	0.000036973	0.0020089	17	22.36842105
hsa05140	Leishmaniasis	74	5.578	2.8684	0.000096299	0.0034882	16	21.62162162
hsa05033	Nicotine addiction	40	3.0151	4.6433	6.02E-07	0.000065465	14	35
hsa05150	Staphylococcus aureus infection	56	4.2212	3.0797	0.00020337	0.0061666	13	23.21428571
UP ACBRI371 CPT								
hsa04115	p53 signaling pathway	72	10,218	88,078	6.71E-03	0.00021880	9	12.5
hsa05210	Apoptosis	136	19,301	41,448	0.00066233	0.042506	8	5.882352941
hsa05222	Colorectal cancer	86	12,205	57,353	0.00020349	0.030728	7	8.139534884
hsa04064	Small cell lung cancer	93	13,199	53,036	0.00033082	0.030728	7	7.52688172
hsa04210	NF-kappa B signaling pathway	95	13,482	51,920	0.00037703	0.030728	7	7.368421053
hsa03018	TNF signaling pathway	110	15,611	44,840	0.00091270	0.042506	7	6.363636364
hsa04668	RNA degradation	79	11,212	53,516	0.00085155	0.042506	6	7.594936709
DOWN ACBRI371 CPT								
hsa05033	Neuroactive ligand-receptor interaction	277	11,126	26,065	0.0000017977	0.000058604	29	10.46931408
hsa04723	Retrograde endocannabinoid signaling	148	59,446	45,420	2.30E-07	3.75E-05	27	18.24324324
hsa04724	Glutamatergic synapse	114	45,789	48,046	5.75E-06	6.24E-04	22	19.29824561
hsa04727	Dopaminergic synapse	131	52,617	38,010	2.20E-03	0.000010239	20	15.26717557
hsa05032	Cell adhesion molecules (CAMs)	144	57,839	34,579	0.0000010483	0.000042717	20	13.88888889
hsa04713	Nicotine addiction	40	16,066	11,826	1.11E-12	3.62E-10	19	47.5

(*continued on next page*)

Table 2 (*continued*)

DOWN LGG - UP GBM								
Gene Set	Description	Size	Expect	Ratio	p value	FDR	LOGFOLD > 2	Enrichment%
hsa04728	GABAergic synapse	88	35,346	50,925	8.42E-05	6.86E-03	18	20.45454545
hsa04514	Morphine addiction	91	36,551	49,246	1.47E-04	9.61E-03	18	19.78021978
hsa05150	Circadian entrainment	96	38,559	46,681	3.56E-04	0.0000019319	18	18.75
hsa04080	Staphylococcus aureus infection	56	22,493	5335	0.0000016155	0.000058518	12	21.42857143
UP U138 CPT	No enriched pathway							
DOWN U138 CPT								
hsa04151	PI3K-Akt signaling pathway	354	73,938	24,345	0.00039531	0.021461	18	5.084745763
hsa04360	Axon guidance	175	36,551	49,246	1.93E-04	0.0000062837	18	10.28571429
hsa04015	Rap1 signaling pathway	206	43,026	37,187	0.0000055099	0.00089812	16	7.766990291
hsa04510	Focal adhesion	199	41,564	36,089	0.000015772	0.0017139	15	7.537688442
hsa04020	Calcium signaling pathway	183	38,222	36,628	0.000025978	0.0021172	14	7.650273224
hsa04540	Gap junction	88	18,380	48,966	0.000083868	0.0054682	9	10.22727273
hsa05146	Amoebiasis	96	20,051	39,899	0.00084677	0.029830	8	8.333333333
hsa04720	Long-term potentiation	67	13,994	50,022	0.00046081	0.021461	7	10.44776119
hsa04971	Gastric acid secretion	75	15,665	44,686	0.00091504	0.029830	7	9.333333333
hsa05143	African trypanosomiasis	35	0.73102	68,397	0.00072934	0.029720	5	14.28571429
UP U251 CPT	No enriched pathway							
DOWN U251 CPT								
hsa05330	Allograft rejection	38	0.020351	98,276	0.00015027	0.020956	2	5.263157895
hsa05332	Graft-versus-host disease	41	0.021957	91,085	0.00017519	0.020956	2	4.87804878
hsa04940	Type I diabetes mellitus	43	0.023029	86,849	0.00019285	0.020956	2	4.651162791
hsa05320	Autoimmune thyroid disease	53	0.028384	70,462	0.00029377	0.023756	2	3.773584906
hsa05416	Viral myocarditis	59	0.031597	63,297	0.00036436	0.023756	2	3.389830508
hsa04612	Antigen processing and presentation	77	0.041237	48.5	0.00062109	0.033746	2	2.597402597

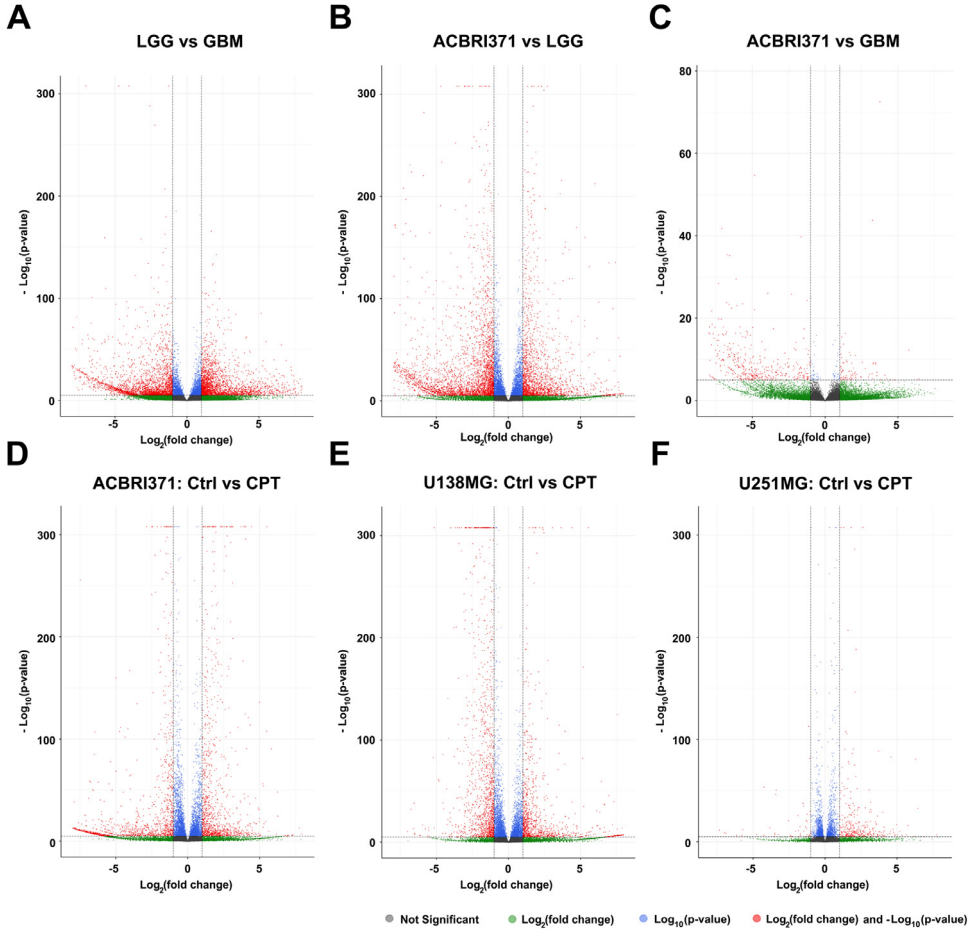

Fig. 2. Volcano plots displaying the degree of altered gene expression among groups of cell lines and treated versus non-treated cells. Volcano plots were produced using the fold change values and p-values generated through the DESeq2 R package analysis to compare the mRNA expression changes between LGG vs GBM (A), ACBRI371 vs LGG (B), ACBRI371 vs GBM (C); and Control cells vs CPT treated cells for: ACBRI371 (D), U138MG (E) and U251MG (F). Blue dots show genes with significant p-value. Green dots show genes with significant fold change. Red dots represent genes with significance in both p-value and fold change.

signaling. While for the comparisons evocative of responses against replication stress, the most enriched pathways were PI3K-Akt signaling, axon guidance, neuroactive ligand-receptor interaction, retrograde endocannabinoid signaling, p53 signaling and apoptosis (Fig. 3). The genes uncovered in each of these pathways are shown in Supplementary Table 1.

2. Experimental Design, Materials and Methods

2.1. Cell culture and treatment

ACBRI371 is a non-tumor human astrocyte cell line that was kindly donated by Prof. Dr. Elza Tiemi Sakamoto Hojo (São Paulo University, Ribeirão Preto, SP, Brazil). HDPC (Human Dental Pulp Cells) is a primary culture of fibroblasts isolated from dental pulp of a 5 years old boy in the

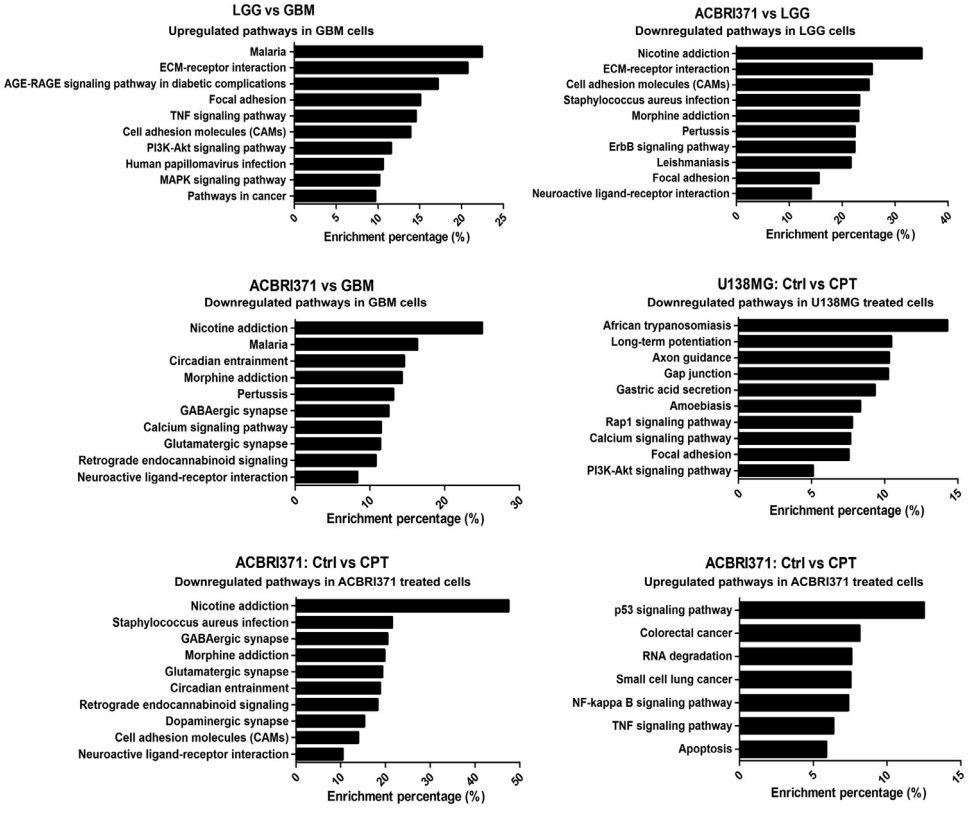

Fig. 3. Enriched KEGG pathways in each collection of genes presenting altered expression in the indicated comparisons. All genes with q-values ≤ 0.0001 (adjusted p-value set to avoid identification of false positive enrichments) and log fold change > 2 or < -2 of each comparison were subjected to pathway analysis by KEGG. Comparisons that revealed enriched pathways are shown. A False Discovery Rate (FDR) ≤ 0.05 were used as threshold to select significant pathways. Graphs were plotted with GraphPad Prism 4.0 software.

laboratory of Dr. C. Costa and Dr. J. Henbling, who gently provided these cells to our laboratory. HDPC were cultivated in standard conditions, with α-MEM (Minimum Essential Medium Eagle) supplemented with 10% of fetal bovine serum and 100 U/mL penicillin and 0.23 mg/mL streptomycin. HDPC was used as an outside cell culture, to be representative of non-brain expression patterns. The cell lines Res186 (grade I), Res286 (grade I), Res259 (grade II) and UW467 (grade III) were all derived from pediatric tumors that were first established by Dr. Michael Bobola (University of Washington, Seattle, WA) and kindly donated to our group by Dr. Fausto Rodriguez (Johns Hopkins University, Baltimore, MD). T98G, U343MG, U87MG, U138MG and U251MG are commercially available GBM cell lines and were obtained from the *American Type Culture Collection*. The non-tumor and GBM cell lines were grown in high-glucose DMEM, while LGG cells were grown in DMEM-F12. All media used were supplemented with 10% fetal bovine serum and 1% penicillin/streptomycin. All cellular stocks were kept in liquid nitrogen before thawed with the appropriate medium. They were all cultured up to a maximum of 75% confluence before and after plating for RNA isolation.

To identify differentially expressed genes associated with tumor progression, we simply cultured the panel of cell lines representative of different grade astrocytoma and compared their RNA-sequencing results. For the replicative stress study, we developed preliminary experiments to choose adequate CPT treatment conditions and the most appropriate cell lines for sequencing,

caring to induce maximum replicative stress yet keeping viable proliferating cells, and selecting one highly CPT-resistant and another CPT-sensitive GBM cell line. In summary, we conducted: (1) MTT dose-response curve to identify the CPT IC_{50} for each GBM cell line and pinpoint the most resistant and sensitive cells, which were U138MG (IC_{50}=3.425 nM) and U251MG (IC_{50}=0.05 nM), respectively; ACBRI371 astrocytes presented an intermediate IC_{50} (1.041 nM) and were also used as control cells (data not shown). (2) H2AX phosphorylation was also accessed in 9 different time points of CPT-treatment at the IC50 for U251MG and U138MG. The peak of H2AX activation was reached around 18 h of treatment, which was then selected as the appropriate time point of analysis (data not shown). Therefore, to evaluate the impact of replicative stress induction on gene expression of GBM cells, we performed RNA-sequencing of U138MG, U251MG and ACBRI371 cell lines treated or not with CPT at the IC_{50} for 18 h.

2.2. RNA isolation and sequencing

Total RNA was isolated with RNeasy Mini Kit (Qiagen) following the manufacturer's instructions. The RNA isolation for each cell line and treatment condition was performed only once (one biological replicate). The density and purity of RNA samples were accessed by 260/280 nm absorbance ratios, measured with NanoDrop spectrophotometer (Thermo Fisher Scientific). The RNA quality was evaluated by electrophoresis in the Bioanalyzer Instrument (Agilent), and samples with an RNA Integrity Number (RIN) $\geq$ 7 were subsequently utilized. For the library construction, 300 ng of high-quality RNA and the TruSeq Stranded Total RNA LT Sample Prep Kit (Illumina Inc.) were applied. Firstly, the RiboZero technology (Illumina Inc.) was used to remove rRNA and preserve only poly(A) and other non-poly(A) transcripts, which were then fragmented. RNA fragments sized between 200 and 500 bp were utilized to generate the sequencing libraries. Clustering was performed in an automated system cBot (Illumina Inc.) and samples were sequenced with TruSeq SBS kit v5, single-read 72 cycles. We have produced two datasets, one sequenced in Genome Analyzer IIx (GAIIx, Illumina Inc.) and another sequenced in NextSeq 500 (Illumina Inc.) (Table 1). These two datasets are technical replicates obtained from the same RNA extraction. All reagents were utilized following the manufacturer's protocols.

2.3. Reads mapping and normalization

The two raw datasets obtained were qualitatively analyzed with FastQC [3]. For each RNA sample analyzed, the GAIIx sequencing generated 8 single-read fastq files obtained from 8 lanes, while the NextSeq 500 dataset yielded 8 paired-end fastq files obtained from 4 lanes. Raw reads quality filtering was accomplished by Trimmomatic [4], removing the Illumina adaptor sequences, low quality bases (phred score quality $>$ 20), and reads shorter than 35 bp. Subsequently, the read error correction was performed by SGA in the preprocess mode (set to -q 25 -f 20 -m 35), followed by index and then correct mode [5]. To evaluate trimming and quality control, the processed reads were inspected by FastQC. Furthermore, the total human transcriptome coverage was assessed for each single-read fastq file or paired-end duo. To achieve a similar coverage distribution for the two different sequencing runs, the GAIIx reads (8 lanes) were randomly grouped in two distinct replicates (A and B), whereas the NextSeq dataset (4 lanes) was kept as four paired-end duos fastq files, each lane representing one replicate (A, B, C and D). This organization was taken forward into the mapping step, in which reads were aligned with the Genome Reference Consortium Human Build 38 (GRCh38) using HISAT2 [6], according to a previously described optimization [7]. Table 1 shows the statistical information for each step.

2.4. Differential gene expression

In order to assess differential gene expression, the number of reads for each transcript was calculated by the HTSeq-count algorithm, settings were operated as -f bam -r pos -s no -a 10 -t

exon -i gene_id -m intersection-nonempty [8]. We conducted a Pearson correlation assessment for all output data using the R function core. Then, the data were directed to the DESeq2 R package for differential expression analysis [9]. Genes were considered differentially expressed when showing an expression change greater than 2-fold, with a *p*-value cutoff of $10e^{-6}$.

2.5. Pathway enrichment analysis

The genes that showed with q-values ≤ 0.0001 (adjusted p-value set to avoid identification of false positive enrichments) and log fold change > 2 or < -2 for each comparison were subjected to pathway analysis using the KEGG database (www.webgestalt.org). A False Discovery Rate (FDR) ≤ 0.05 was used as a threshold to select significant pathways. Pathway enrichment charts were plotted with GraphPad Prism 4.0 software.

Ethics Statement

The authors declare that this study did not involve any human or animal subjects.

CRediT Author Statement

Juliana de Sousa: Investigation, Validation, Writing - Original draft preparation, Writing - Reviewing & Editing; Patrick da Silva: Formal analysis, Validation, Data curation, Visualization, Writing - Original draft; Rodolfo Serafim: Formal analysis, Data curation, Visualization; Ricardo Nociti: Formal analysis; Cristiano Moreira: Supervision; Wilson Silva Jr: Supervision, Resources; Valeria Valente: Conceptualization, Resources, Supervision and Data analysis, Project administration, Funding acquisition, Writing - Original draft and Reviewing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships, which have, or could be perceived to have, influenced the work reported in this article.

Acknowledgments

This study was supported by Fundação de Amparo à Pesquisa do Estado de São Paulo - FAPESP (grant #2013/13465-1; grant #2018/05018-9), Programa de Apoio ao Desenvolvimento Científico - PADC from Faculty of Pharmaceutical Sciences of Araraquara and by Center for Cell-Based Therapy (CEPID/FAPESP; grant #2013/08135-2). We also thank CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) and CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) for fellowships granted to students. The authors acknowledge Kamila Peronni for all technical support in libraries construction and sequencing.

Supplementary Materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.dib.2020.106643.

References

[1] D.N. Louis, A. Perry, G. Reifenberger, A. von Deimling, D. Figarella-Branger, W.K. Cavenee, et al., The 2016 World Health Organization classification of tumors of the central nervous system: a summary, Acta Neuropathol. 131 (2016) 803–820 10.1007/s00401-016-1545-1.

[2] J.F. de Sousa, R.B. Serafim, L.M. de Freitas, C.R. Fontana, V. Valente, DNA repair genes in astrocytoma tumorigenesis, progression and therapy resistance, Genet. Mol. Biol. (2020) 43 10.1590/1678-4685-gmb-2019-0066.

[3] S. Andrews, FastQC: a quality control tool for high throughput sequence data, FastQC A Quality Control Tool High Throughput Sequence Data. (2010) Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ .

[4] A.M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data, Bioinformatics 30 (2014) 2114–2120 https://doi.org/10.1093/bioinformatics/btu170.

[5] J.T. Simpson, R. Durbin, Efficient de novo assembly of large genomes using compressed data structures, Genome Res. 22 (2012) 549–556 https://doi.org/10.1101/gr.126953.111.

[6] D. Kim, J.M. Paggi, C. Park, C. Bennett, S.L. Salzberg, Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype, Nat. Biotechnol. 37 (2019) 907–915 https://doi.org/10.1038/s41587-019-0201-4.

[7] G. Baruzzo, K.E. Hayer, E.J. Kim, B. Di Camillo, G.A. FitzGerald, G.R. Grant, Simulation-based comprehensive benchmarking of RNA-seq aligners, Nat. Methods. 14 (2017) 135–139 https://doi.org/10.1038/nmeth.4106.

[8] S. Anders, P.T. Pyl, W. Huber, HTSeq–a Python framework to work with high-throughput sequencing data, Bioinformatics 31 (2015) 166–169 https://doi.org/10.1093/bioinformatics/btu638.

[9] M.I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2, Genome Biol. 15 (2014) 550 https://doi.org/10.1186/s13059-014-0550-8.