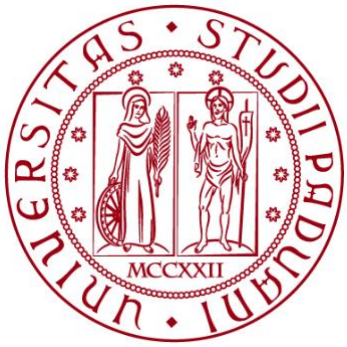
UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI BIOLOGIA

Corso di Laurea magistrale in Molecular Biology



TESI DI LAUREA

Integrative pathway analysis of gynecological tumors characterized by different chromosomal instability patterns

Relatore: Prof.ssa Chiara Romualdi
Dipartimento di Biologia

Laureanda: Anna Bortolato

ANNO ACCADEMICO 2022/2023

Abstract

Gynecological tumors include four tumor types from TCGA: ovarian cancer (OV), cervical cancer (CESC), endometrial cancer (UCEC) and the rare uterine carcinosarcoma (UCS). Breast cancer (BRCA) can also be included among gynecological tumors, since it shares the same embryonic origin and the influence of female hormones. In order to shed light on the common molecular features characterizing the five tumors, molecular profiles from patients with different patterns of chromosomal instability (CIN), underlying different mechanisms of dysregulation (signatures), were compared at both expression and methylation level. A pathway analysis was performed using SourceSet software, whose topological approach allows to discriminate genes that are the primary source of dysregulation from those that are indirectly affected. Two signatures were investigated: CX1 and CX3. CX1 is characterized by defective mitotic spindle checkpoint, resulting in incorrect chromosome segregation, while CX3 shows replication stress, leading to double strand breaks that are not properly corrected by homologous recombination and result in structural aberrations. Results revealed that OV, UCEC and BRCA tend to have similar expression and methylation profiles, while UCS and CESC are the most divergent tumors. Primary genes are more uniformly detected across tumor types compared to secondary genes, reflecting the common origin of perturbation generating the observed pattern of CIN and the different molecular profiles characterizing different tissues, respectively.

Contents

1 Introduction

1.1 Gynecological tumors

Gynecological tumors have an estimated incidence of almost 400,000 cases and more than 70,000 deaths among United States female population in 2023 (Table 1).

Tumor type	Incidence	Mortality
Breast	297,790	43,170
Uterine cervix	13,960	4,310
Uterine corpus	66,200	13,030
Ovary	19,710	13,270
Total	397,660	73,780

Table 1: Estimated new cases and deaths for gynecological cancers among USA female population updated to 2023[1].

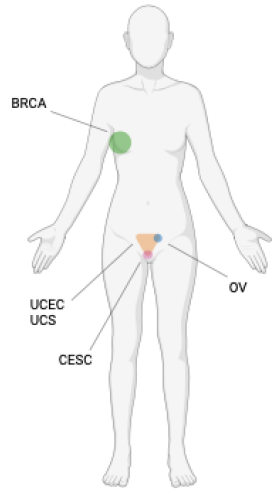


Figure 1: Location of gynecological tumors in female body.

Gynecological tumors (Figure 1) include four tumor types from TCGA: high-grade serous ovarian cystadenocarcinoma (OV), cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), uterine corpus endometrial carcinoma (UCEC), uterine carcinosarcoma (UCS). A fifth tumor type can be included in the list of gynecological tumors: invasive breast carcinoma (BRCA), as it shares important characteristics with proper gynecological cancers: they arise from the same embryonic tissues, they are all influenced by female hormones, they are treated by the same medical specialty, gynecologic oncology. Despite recent clinical advances, molecular characteristics of these tumors are not completely uncovered. In the following sections a brief description of the five gynecological cancer types is proposed, focusing on their similarities and unique molecular features, as they were derived from multi-platform studies by TCGA Research Network.

1.1.1 Ovarian cancer

Ovarian cancer is the fifth most mortal cancer type among women in United States (5%), most deaths (∼70%) are imputable to the most aggressive high-grade serous ovarian cancer (HGSOC), characterized by P53 mutations and genomic instability due to defects in DNA repair pathways. Familial cases are

found due to germline mutations in homologous recombination-mediated repair genes BRCA1 and BRCA2 (10-20% cases). The principal factor influencing the elevated mortality of HGSOC patients is the inability to diagnose the disease at early stages, due to a lack of screening methods and symptoms. Integrated genomic analyses on HGSOC patients[2] found TP53 is ubiquitously mutated across tumors (96%), with only a few other genes showing recurrent mutations, including BRCA1 and BRCA2. Genome instability is profound, with amplifications on MYC and CCNE1. Homologous recombination pathway of DNA repair is defective in 51% of cases, with genes affected BRCA1/2, PTEN, RAD51C, ATM, ATR and Fanconi anaemia genes. Other frequently altered pathways include: RB1, PI3K/Ras, NOTCH, FOXM1. Treatments currently available include PARP inhibitors, anti-angiogenic factors, platinum-based chemotherapy.

1.1.2 Endometrial cancer

Endometrial cancer arises from the inner epithelial lining of the uterus. It is the third most common cancer type affecting women in the United States, with incidence rates constantly increasing since the mid-1990s of 2% cases per year among young women. Most patients present with low-grade, early-stage disease. Among this cancer type, two groups have been distinguished: endometrioid tumors, linked to estrogen excess, obesity, receptor-positivity and favorable outcome, and serous tumors, more frequent in older, non-obese women, associated with a worse prognosis. A study from TCGA Research Network[3] identified four molecular subgroups. The first subgroup is characterized by extensive copy number changes and includes high-grade aggressive cancers, mainly from the serous histological type. TP53 is mutated in most of the tumors, with frequent mutations also in FBXW7 and PPP2R1A, while CCNE1 and ERBB2 are frequently amplified. Uterine serous carcinomas share many molecular features with HGSOC and basal-like breast carcinoma. Another subgroup is characterized by microsatellite instability. Commonly altered genes are PTEN, ARID1A, PIK3CA, RPL22, MLH1. The third subgroup shows recurrent mutations in the exonuclease domain of POLE, polymerase-ϵ, with consequent increased replication error frequency, leading to a great mutational burden. This subgroup presents the most favorable outcome, thanks to the high lymphocytic infiltration. The fourth subgroup presents a low amount of copy number alterations and low mutational burden. It contains low grade tumors.

1.1.3 Breast cancer

Breast cancer is the most common cancer among women worldwide. Clinically it is categorized into three therapeutic groups: the estrogen receptor

positive, that can be treated with endocrine therapy, the HER2 amplified group, treated by targeting HER2, and the triple negative cancers (TNBC), with only chemotherapy options. When combining information from different platforms[4], a large heterogeneity of molecular features is identified for this cancer type, with only three genes, TP53, PIK3CA and GATA3 showing somatic mutations at $> 10\%$ incidence. TNBC and HGSOC comparison indicated several molecular commonalities, such as TP53, RB1, BRCA loss and MYC amplification.

1.1.4 Cervical cancer

95% cases of cervical cancer are caused by persistent infections by HPV. Cervical cancer incidence is clearly decreasing since 1970s worldwide, more than any other gynecological tumor, because of the diffusion of HPV screenings. Vaccination campaigns and novel strategies for screening are now available in developed countries, producing disparities in the incidence rates of the tumor in developed and under-developed countries. PI3K-MAPK and TGFβ signaling pathways are frequently altered. APOBEC mutational signature correlates with the total number of mutations per sample, suggesting a role of cytidine deaminases and mRNA editing in cervical carcinogenesis. A subgroup of endometrial-like cervical cancers was identified by multi-platform studies[5], composed predominantly of HPV-negative tumors, with high frequencies of KRAS, ARID1A and PTEN mutations.

1.1.5 Uterine carcinosarcoma

UCS tumors are biphasic carcinomas, showing morphological components of both epithelial and mesenchymal cell types. These tumors are very rare, they represent 5% of the total number of uterine cancers and are associated with poor prognosis (15% of deaths for uterine malignancies). They arise from epithelial cells of the uterus that undergo differentiation into mesenchymal cells. Frequently mutated genes were also found in endometrial cancer, such as TP53, FBXW7, PPP2R1A and genes from the PI3K pathway. Transcriptome analyses revealed a strong epithelial-to-mesenchymal transition gene signature, with altered expression of E-cadherin, and SNAI1/2 and ZEB1/2 regulation by members of miR-200 family. These miRNAs are in turn regulated by methylation at their promoters[6].

1.2 Chromosomal instability signatures

One of the most recognizable hallmarks of cancer is chromosomal instability (CIN), deriving from the accumulation of genomic alterations that can involve either a gain or loss of whole chromosomes or structural aberrations,

ranging from small-scale insertions or deletions to large DNA rearrangements. CIN is responsible for the intratumoral heterogeneity that drives phenotypic adaptation during tumor evolution and it is often involved in anticancer drug resistance.

Aberrant chromosome segregation is often responsible for CIN and may derive primarily from mitotic defects, such as altered microtubule-spindle dynamics and defects affecting the mitotic checkpoint or sister-chromatid cohesion. Multiple centrosomes are often observed in cancer cells, causing defects on microtubule-kinetochore attachment, that lead to inactivation of mitotic checkpoint. Genome doubling, or tetraploidization, arising from failed cell division or endoreplication, is also frequent across several cancers and responsible for an increased risk of chromosome missegregation. Pre-mitotic defects, such as replication stress, can generate chromosome fusions, that lead to formation of acentric or dicentric chromosomes and consequently unequal distribution of genetic material, even without defects in chromosome segregation machinery. Furthermore, aberrantly segregated chromosomes may involve the formation of isolated DNA surrounded by nuclear envelope, i.e., micronuclei, that are associated to disruption of nuclear envelope with consequent exposure of nuclear DNA to reactive oxygen species and cytoplasmic enzymes, such as RNA editing enzymes from the APOBEC family, resulting in the accumulation of mutations and further structural defects.
For the propagation of CIN across cancer cells the disruption of DNA damage response is fundamental: P53 pathway is inactivated, by mutations on TP53 or indirectly by affecting other genes of the pathway. Immune evasion is also responsible for the proliferation of chromosomally unstable cancer cells, e.g., by decreasing expression of genes involved in adaptive immunity or cytotoxicity mediated by CD8+ T cells and natural killer cells.
Understanding the molecular heterogeneity of CIN and its connection with differential clinical outcomes may be exploited to develop new approaches for cancer treatment and diagnosis, especially in the perspective of personalized medicine[7].

A study by Drews et al.[8] published in 2022 investigated the different features of CIN in a pan-cancer analysis, linking a specific aetiology to a particular profile of CIN. Copy number profiles across 33 TCGA tumor types were collected and samples with detectable CIN were selected. Distributions of fundamental copy number features (e.g., breakpoint counts per chromosome arm, copy number change between a segment and the neighboring, segment length) representing different causes of CIN were computed. Using a mixture modeling approach, they derived 17 pan-cancer copy number signatures (Figure 2).
To determine the putative aetiology for each signature, they considered both patterns of copy number change and signature association with known cancer

driver mutations. A confidence score was also assigned to each signature aetiology based on the quality and extent of supporting data. Finally, they were able to quantify the activity of each signature across patients.
Considering only gynecologic cancer patients, signature 1 (CX1) and 3 (CX3) are the most active (Figure 3).

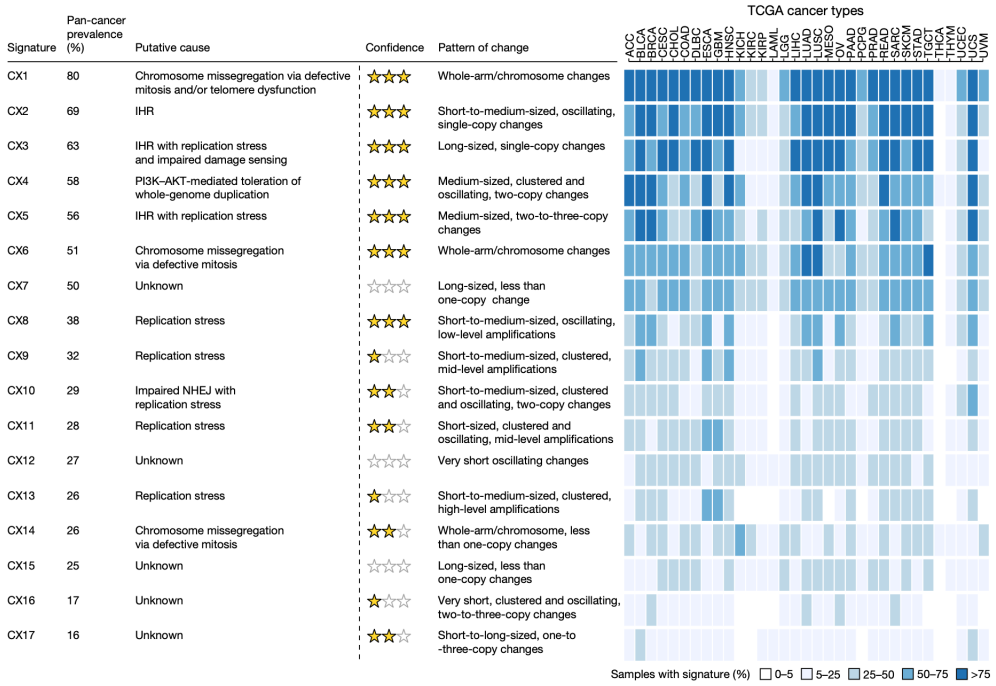


Figure 2: Heatmap showing the prevalence of each signature across 33 TCGA tumor types. Each signature is described with its putative cause, confidence score and pattern of CNV[8].

1.2.1 Signature 1

Signature 1 is characterized by whole-arm or whole-chromosome changes, suggesting as a putative cause chromosome missegregation via defective mitosis and telomere dysfunction; indeed, it is negatively correlated with telomerase expression and telomere length. The signature has higher activity in mutated CIC, VHL and PBRM1 carriers.

1.2.2 Signature 3

Signature 3 exhibits long-sized, single-copy changes, patterns that are associated to impaired homologous recombination. Activity of this signature is increased in patients with germline mutation of BRCA1 and BRCA2 and methylated RAD51C. Replication stress is also involved in the aetiology of

the signature (via amplification of MAPK1, PPP2R1A, U2AF1). In addition, key nucleotide excision repair genes are downregulated, as well as TP53, suggesting impaired damage sensing.

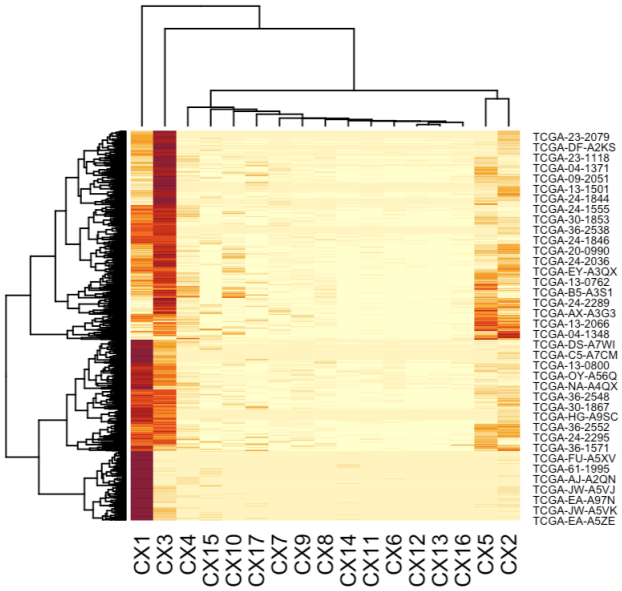


Figure 3: Heatmap showing signature levels for each gynecologic tumor patient. Low values are yellow, high values are red.

1.3 DNA methylation in cancer

Chromosomal instability is known to influence gene expression in cancer cells, but other factors are frequently involved. DNA methylation, histone modifications and chromatin remodeling are epigenetic factors able to influence gene expression without requiring any modification in DNA sequence. DNA methylation is contributing to the final pattern of gene expression that we can observe in cancer cells, where these changes are then inherited throughout cell divisions thanks to DNA methylation maintenance machinery.
DNA methylation occurs mainly on cytosine, forming 5-methylcytosine. This modification is observed with high frequency on CpG islands, regions rich in the dinucleotide CG in 5'-3' direction. CpG islands are DNA sequences roughly 1000 bp long with a GC content $> 50\%$. About 70% of human genes contain CpG islands in their promoter.
Most often, DNA methylation acts by silencing genes via hypermethylation on CpG islands in the promoter region of genes, but there are different possible mechanisms in which it can regulate gene expression, for example, by affecting expression of miRNAs. Based on this knowledge, DNA methylation is expected to be responsible for transcriptional gene silencing more often than

DNA sequence mutation.

The silencing of DNA repair genes via hypermethylation may be an early step in carcinogenesis: deficiency of DNA repair genes leads to accumulation of DNA damages that give rise to cancer. miRNAs are involved in gene silencing by targeting mRNAs of protein coding genes; differential methylation on miRNA promoters may influence their expression, and thus indirectly affect gene expression. DNA hypermethylation is observed in the promoter regions of tumor suppressor genes, involved in several cancer types, such as RB1, CDKN2A, CDKN2B (regulating cell cycle), DAPK1 (involved in apoptotic signaling), and cell adhesion molecules (CDH1, CDH13). Hypermethylated promoters act by recruiting transcriptional repressors and histone-modifying enzymes, while inhibiting the binding of transcription factors to DNA.
In contrast, oncogenes are often associated with hypomethylation of cancer-specific CpG islands. However, in cancer cells genome-wide hypomethylation is observed, that can be responsible for chromosomal instability, derepression of imprinted genes and retrotransposons, as well as aberrant gene expression[9].

1.4 Pathway enrichment analysis

To study the impact of DNA methylation on gene expression, high-throughput experiments are performed, such as RNA sequencing and DNA methylation arrays. They return a value for each single gene or CpG island that is used to quantify its level of expression or methylation. Generally, an RNAseq experiment quantifies more than 20,000 genes across a large number of samples. Similarly, Illumina arrays for genome-wide methylation studies can contain either 27,000 or 450,000 probes for known CpG islands. Indeed, the challenge of working with high amounts of data is to analyze them and extract useful information. When comparing two conditions, we can search for groups of related genes that are significantly altered, in order to reduce the dimension of the results and make them more informative: this is the main purpose of a pathway enrichment analysis.

Different methods of enrichment analysis have been developed[10]. The first method for enrichment analysis, the most simple one, is over- representation analysis: starting from a list of selected differentially expressed genes (DEGs), it detects as significantly enriched pathways those containing a greater proportion of DEGs than expected by chance. With this approach a pathway is represented as a list of genes, without any knowledge about the interactions between them.

1.4.1 Topological pathway analysis

Indeed, a pathway can be described not only as a group of genes involved in the same biological theme, but also as a set of genes and the pairwise interactions between them: it can be represented as a graph with nodes as genes and edges as their biochemical interactions. The character of these interactions can be directed, if the presence of a gene is affecting in a specific direction another gene, or undirected.
With this definition, it is possible to perform topological pathway analysis: for each component of the pathway, this approach takes into account both its differential expression and the effect of its dysregulation on interacting genes, i.e., how the entire pathway would be affected by a specific gene dysregulation. The final significance of the pathway is corrected by multiple-testing error correction.
Topological pathway analysis is the most recent generation of enrichment analysis methods; it allows to highlight the most interesting pathways that are clearly altered between two conditions by testing simultaneously gene expression level and pathway structure.

1.4.2 SourceSet

Multiple methods are now available for topological pathway analysis. *SourceSet* is one of few methods that is able to distinguish genes that are the source of perturbation from genes that merely respond to the dysregulation, as the effect of network propagation[11]. Primary genes detected by *SourceSet* show few overlaps with top ranked differentially expressed genes, suggesting that a slightly altered gene can deeply affect several downstream genes and be the cause of perturbation. This novel approach is fundamental to prioritize the effect of biological perturbations in network medicine, allowing a better understanding of drug treatments and diseases. It works comparing gene profiles in control and in perturbed conditions and detects differences in both the mean and the covariance parameters. Briefly, this is the workflow of the algorithm.

- Pathways are treated as graphs and they are converted into decomposable undirected graphs.
- After decomposition each graph is made of cliques and separators (Figure 4).
- Marginal test statistics are calculated for each clique and separator. Conditional test statistics is also calculated as the difference between clique and separator marginal test statistics. Both equality of mean and covariance parameters are tested for each component of the clique (gene) by likelihood ratio test. Multiple testing error correction is applied to estimate the significance of the clique.

- The union of cliques found to be significantly dysregulated is computed: this is the secondary set.
- The source set is the intersection of cliques found dysregulated across decompositions.

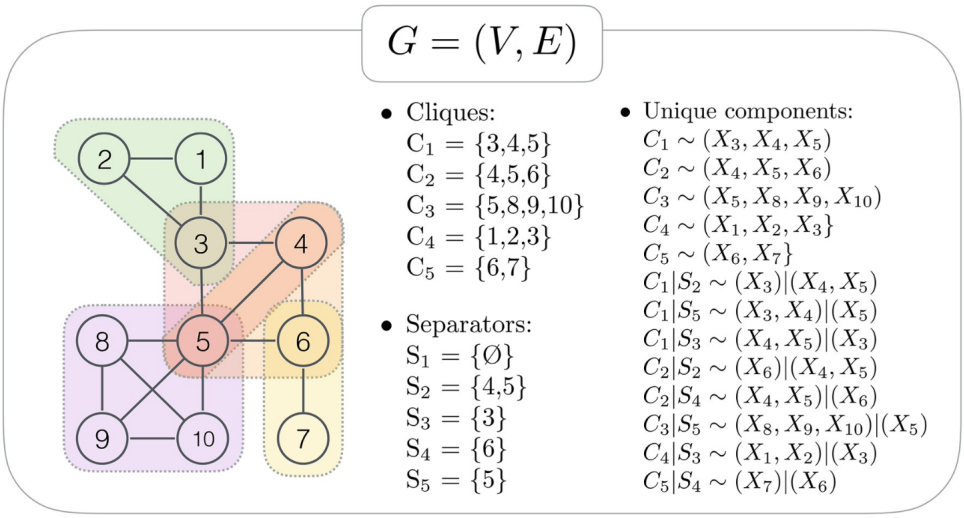


Figure 4: Example of cliques and separators from the authors paper[11].

1.5 Aim of the project

Gynecological tumors have been widely analyzed separately, but few studies exist that aim to investigate their common molecular features. One of the most complete studies that compared gynecologic tumors at molecular level was performed by TCGA Research Network[12]. They were able to detect gynecologic tumor-specific molecular features that differ in frequency compared to 28 non-gynecologic TCGA tumor types, including amplifications, deletions and mutations; with these results the authors were able to identify prognostic molecular subtypes that could be interesting also as therapeutic targets in a cross-cancer approach.

Focusing on the two signatures illustrated above, in this study we decided to investigate and characterize gynecological cancer patients with different signature activity, in order to find which pathways are dysregulated as a result of a specific pattern of chromosomal instability, and thus shed light to the molecular commonalities of gynecological cancers.
For this purpose, a pathway analysis is performed, comparing molecular profiles of gynecologic tumor patients exhibiting high activity of a signature with

patients where the same signature is consistently inactivated. To have a more global overview of the different profiles of patients, the analysis is performed at both expression and methylation level. Indeed, expression levels of a gene are often regulated by methylation, frequently in the promoter region. The analysis will reveal more in detail which mechanisms are involved in the aetiology and effects of the different signatures, by returning affected pathways and genes. Results will also allow the recognition of similarities and differences between the five tumors and the effects of methylation perturbations on expression profiles (Figure 5).
These findings would be useful for prognostic and therapeutic purpose: it is possible to associate a specific signature to a better or worse clinical outcome and resistance to specific treatment. Moreover, based on the signature activation, specific treatments could be developed for single cancer and also cross-cancer therapies, considering the similarities between the five gynecologic cancers.

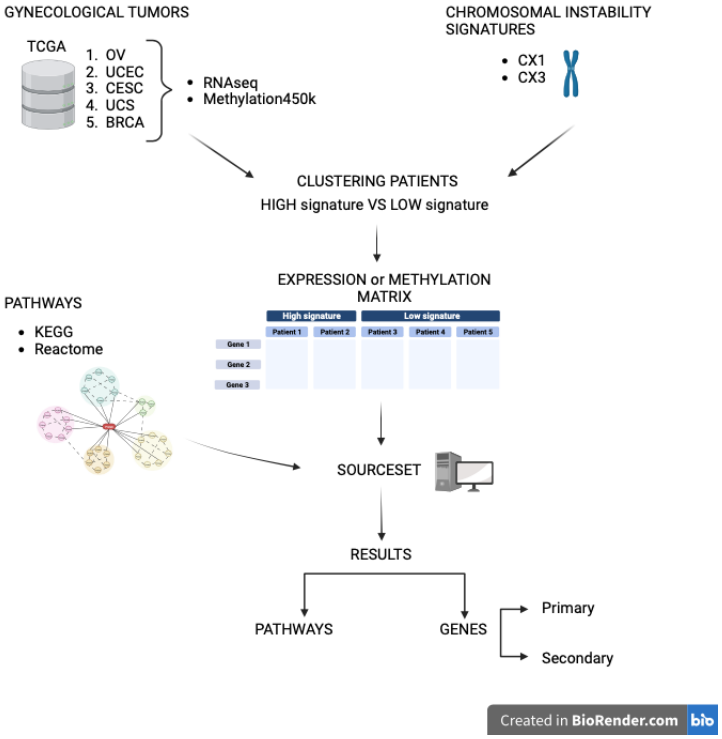


Figure 5: Workflow of the analysis.

2 Materials and Methods

2.1 Data download

Experimental assays for gene expression and methylation were retrieved from TCGA data portal by *curatedTCGAData* R package (2.0.1 data version). Disease codes for gynecologic tumor types were BRCA, OV, UCEC, UCS, CESC. Selected data types were RNASeq2GeneNorm and Methyl*. Male BRCA patients were discarded.

RNAseq data contain upper-quartile normalized RSEM TPM counts (UCEC has two RNAseq assays available, obtained from different Illumina sequencing platform: the Genome Analyzer data was preferred, since it contained more samples).

Methylation data are composed of β-values obtained from Illumina BeadChip arrays, where DNA fragments are applied after bisulfite conversion. DNA molecules hybridize with CpG locus-specific oligomers, linked to two different bead types, one for the methylated and one for the unmethylated state; single-base extension using a labeled nucleotide follows the hybridization. Beta-values are continuous variables spanning from 0 to 1, calculated as the ratio of the fluorescence intensity of the methylated bead type with respect to the combined locus intensity[1]. For some cancer types assays from Illumina HumanMethylation 27k and Infinium HumanMethylation 450k BeadChip were both available: 450k assay was preferred (only for OV 27k assay was used since it contained more samples).

The matrix defining the levels of each signature across patients was downloaded from the GitHub page of the authors of the study on chromosomal instability[8]. The table of the metadata to map each patient to a specific cancer type was also downloaded from the page.

2.2 Data preparation

2.2.1 RNAseq

RNAseq counts were log-transformed. Genes that were not expressed in more than 50% samples were discarded. Gene names were converted into Entrez IDs.

[1]Comprehensive DNA Methylation Analysis on the Illumina® Infinium® Assay Platform, Contributed by Daniel J. Weisenberger, David Van Den Berg, Fei Pan, Benjamin P. Berman, and Peter W. Laird, University of Southern California, Keck School of Medicine, USC/Norris Comprehensive Cancer Center, Los Angeles, CA 90033

2.2.2 Methylation beta-values

Mean β-value is calculated from CpG islands falling in the same gene (information retrieved from metadata). β-values were transformed into quantiles of standard normal distribution. Genes with NA values across all samples were discarded. Gene symbols were converted to Entrez IDs. The remaining NA values were imputed by K-nearest neighbors algorithm (*impute.knn* function from *impute* R package).

2.2.3 Pathways

Lists of pathways from Reactome and KEGG databases were retrieved by *graphite* R package. Gene names were converted into Entrez IDs. Pathways topology was built with *pathwayGraph* function, in order to obtain graphNEL objects. Graphs with more than 300 or less than 5 nodes were discarded. The final number of pathways on which the analysis was performed is 306 for KEGG and 1712 for Reactome.

2.2.4 Clustering

Patients for each tumor were clustered according to the level of each signature activation preferentially by *mclust* R package with a Gaussian mixture modeling approach. When multiple groups were detected, groups with high signature or low signature were joined together, in order to obtain two groups. For BRCA, UCEC, UCS CX3 it was not possible to identify two groups of patients with *mclust*: in this case hierarchical clustering was used.

2.3 SourceSet analysis

SourceSet analysis was run with permute and shrink parameters set to TRUE and seed=111 (version 0.1.5). Input data were the list of graphs, the expression or methylation matrix with genes as columns and samples as rows and a vector defining the lengths of each group of patients.
The analysis is run for each tumor and each signature at both expression and methylation level for a total of 20 analyses (Table 2).

Signature	Omic	Tumor	N Low	N High	Tot
CX1	Expression	BRCA	345	328	673
		UCEC	71	53	124
		OV	200	73	273
		CESC	59	155	214
		UCS	40	14	54
	Methylation	BRCA	213	245	458
		UCEC	110	71	181
		OV	376	140	516
		CESC	60	157	217
		UCS	40	14	54
CX3	Expression	BRCA	556	117	673
		UCEC	106	18	124
		OV	172	101	273
		CESC	150	64	214
		UCS	18	36	54
	Methylation	BRCA	387	71	458
		UCEC	158	23	181
		OV	313	203	516
		CESC	151	66	217
		UCS	18	36	54
Tot		20 analyses			

Table 2: *SourceSet* analyses with size for each class of patients.

2.4 Processing of the results

Results from each analysis were first processed independently. The output of *SourceSet* is a list of lists, one for each input graph. For each pathway a primary and a secondary set of genes are detected. Primary and secondary genes were extracted from each graph and merged together in order to have two sets of genes for each analysis. An enrichment analysis on Gene Ontology is performed for both primary and secondary genes with the *ClusterProfiler* package. For each input graph and for each node of input graphs, parameters were calculated with the *infoSource* function from *SourceSet*.

Moreover, since *SourceSet* was not able to distinguish whether a gene was activated or inactivated is a specific class of patients, an additional parameter was calculated for each gene, i.e., the logFC, comparing expression or methylation level of the gene from patients with high CX1 against patients with high CX3, or vice versa (based on which signature was considered for clustering patients). To avoid overlappings in the two groups of patients, only the third quartile of patients was used to calculate the fold-change. The same measure is also included in graphs and GO enrichment results, where it is calculated as the median logFC of primary genes belonging to the pathway.
Plots were generated using *ggplot2* and *UpSetR* packages.

2.4.1 Example of results

Using the *infoSource* function, several variables are calculated for each graph and node of the graphs. The logFC variable is calculated separately and appended to the tables.

	n.primary	n.secondary	n.graph	n.cluster	primary. impact	total. impact	adj. pvalue	logFC (median)
Phosphorylation of Emi1	6	0	6	1	1	1	0.00	-0.32
Defective CHST6 causes MCDC1	7	0	7	1	1	1	0.00	0.32
Interleukin-18 signaling	5	0	5	2	1	1	0.01	0.24
Condensation of Prometaphase Chromosomes	11	0	11	1	1	1	0.00	-0.24
CDC6 association with the ORC:origin complex	2	0	2	1	1	1	0.00	-0.23
Unwinding of DNA	11	0	11	1	1	1	0.00	-0.22
E2F-enabled inhibition of pre-replication complex formation	3	0	3	1	1	1	0.03	-0.18

(a) Top 7 Reactome pathways, sorted by primary.impact and absolute value of logFC.

	mapped.name	n.primary	n.secondary	n.graph	specificity	primary.impact	total.impact	score	relevance	logFC
5295	PIK3R1	36	53	126	0.07	0.29	0.71	2.41	0.02	0.63
472	ATM	26	5	52	0.03	0.50	0.60	1.56	0.02	0.07
4683	NBN	26	3	35	0.02	0.74	0.83	2.21	0.02	-0.01
10111	RAD50	25	3	35	0.02	0.71	0.80	2.21	0.01	0.06
5436	POLR2G	24	5	60	0.04	0.40	0.48	1.87	0.01	-0.39
5440	POLR2K	24	5	79	0.05	0.30	0.37	1.84	0.01	-0.28
5433	POLR2D	24	5	60	0.04	0.40	0.48	1.87	0.01	-0.25

(b) Top 7 dysregulated genes from Reactome pathways, sorted by relevance and absolute value of logFC.

Table 3: Example of the output generated by *infoSource* function on the results for the analysis on signature 1 of OV expression data, with the addition of the logFC column.

Variables calculated for each graph (Table 3a) include:

- number of primary and secondary genes detected for the pathway (**n.primary**, **n.secondary**),
- total number of nodes composing the graph (**n.graph**),
- number of connected components of the graph (**n.cluster**),
- proportion of primary genes with respect to the total size of the graph (**primary.impact**),
- proportion of all dysregulated genes over the total graph size (**total.impact**),
- adjusted p-value for the hypothesis of equality of the two distributions associated to the graph (**adj.pvalue**),

- median logFC of primary genes belonging to the pathway, calculated between patients with high CX1 and patients with high CX3 (**logFC**).

Variables calculated for each node of the graph (Table 3b) include:

- number of graphs whose primary set contains the gene (**n.primary**),
- number of graphs whose secondary set contains the gene (**n.secondary**),
- number of graphs containing the gene (**n.graph**),
- proportion of graphs containing the gene over the total number of input graphs (**specificity**),
- percentage of graphs whose source set contains the gene over the number of graphs in which the gene appears (**primary.impact**),
- percentage of graphs in which the gene was found dysregulated with respect to the number of input graphs containing the gene (**total.impact**),
- combination of p-values from all graphs containing the variable, ranging from 0 to $+\infty$, where higher values indicate higher significance (**score**),
- percentage of input graphs that contain the gene in their source set, with respect to the total number of input graphs (**relevance**),
- differential level of expression between patients with high CX1 and patients with high CX3 (**logFC**).

Significant pathways were selected using a threshold of adjusted p-value (by default set to 0.05); when the number of available samples was too small, the threshold was set on higher values in order to keep a reasonable amount of significant pathways. Selected pathways were sorted according to their primary impact and absolute logFC. Pathways with a primary impact equal to 0 were discarded.
Genes were separated into primary and secondary genes and sorted according to their relevance or total impact, respectively. Genes with the same relevance or total impact were further sorted according to their absolute value of logFC.

3 Results

The amount of pathways detected as significant (Tables 4a, 4b) and the size of the primary and secondary set of genes (Tables 5a, 5b) may vary a lot among analyses, mainly due to the different availability of samples across tumors, affecting the sensitivity of the algorithm to smaller perturbations. Breast cancer with the highest number of samples (expression = 673, methylation = 458) shows the largest amount of primarily dysregulated genes and significant pathways detected, for both KEGG and Reactome pathways, while UCS with the least amount of samples (54) is most frequently the tumor with the smallest estimated source set and the least amount of significant pathways.

Reactome

		BRCA	UCEC	OV	CESC	UCS
CX1	EXPRESSION	1709	1401	304	46	0
	METHYLATION	1691	1273	192	29	3 (0.1)
CX3	EXPRESSION	1703	160 (0.1)	737	81	2 (0.1)
	METHYLATION	1573	1 (0.1)	366	14	0

(a) Significant Reactome pathways.

KEGG

		BRCA	UCEC	OV	CESC	UCS
CX1	EXPRESSION	306	268	61	17 (0.1)	0
	METHYLATION	306	266	62	0	0
CX3	EXPRESSION	306	0	141	0	0
	METHYLATION	301	0	73	0	0

(b) Significant KEGG pathways.

Table 4: Number of significant pathways detected in each analysis. The threshold of adjusted p-value used to filter the results for significant pathways is specified inside parentheses (when not present it was the default 0.05).

Reactome

			BRCA	UCEC	OV	CESC	UCS
CX1	EXPRESSION	Primary	7034	4166	978	1114	164
		Secondary	704	2057	987	837	462
	METHYLATION	Primary	6522	4058	811	108	146
		Secondary	1222	1833	917	490	362
CX3	EXPRESSION	Primary	6667	1225	2286	578	103
		Secondary	989	969	1989	665	264
	METHYLATION	Primary	5469	129	1131	614	37
		Secondary	1624	459	1151	1076	357

(a) Primary and secondary genes detected for Reactome pathways.

KEGG

			BRCA	UCEC	OV	CESC	UCS
CX1	EXPRESSION	Primary	4311	1658	259	67	33
		Secondary	341	1696	664	302	131
	METHYLATION	Primary	3959	1808	324	15	29
		Secondary	674	1465	413	108	105
CX3	EXPRESSION	Primary	4054	60	521	9	0
		Secondary	525	248	1136	237	57
	METHYLATION	Primary	2940	0	270	19	1
		Secondary	1286	76	507	191	70

(b) Primary and secondary genes detected for KEGG pathways.

Table 5: Number of primary and secondary genes detected for each analysis.

The peculiarity of Reactome database is the hierarchical organization: different pathways represent the same process with more or less details, going from very specific and small pathways to very large and general ones. Indeed, they are more focused on the specific portions of a biological process, even taking into account the action of specific molecules on them and how their dysregulations are implicated in specific diseases. On the other side, KEGG pathways are way more general, depicting complex biological processes in a single pathway. This different approach in building pathways leads to an average number of nodes that differs between the two pathway databases, with KEGG pathways having a higher average number of nodes ($\sim$79) compared to Reactome pathways ($\sim$44).
For simplicity, the plots representing the results in the next sections will be relative to *SourceSet* analyses on Reactome pathways. Results relative to KEGG pathways are included in the Appendix section.

3.1 Expression

3.1.1 Pathways

Results obtained from expression data are compared considering only a subset of the large amount of significant pathways detected by the software, consisting of the top 10 pathways for primary impact and absolute value of logFC for each analysis (Figure 6).

Pathways related to chromosome segregation including centrosome maturation, recruitment of NuMA to mitotic centrosomes anchoring of the basal body to the plasma membrane, deposition of CENPA containing nucleosomes at the centromere and condensation of prometaphase chromosomes tend to be inactivated in CX1. Moreover, the inhibition of the anaphase-promoting complex (APC/C) by mitotic spindle checkpoint components is affected in

Figure 6: Heatmap showing the top 10 Reactome pathways for primary impact and absolute value of logFC detected across analyses at expression level. The color of cells corresponds to the median logFC for the pathway, while the number inside cells is the primary impact. Grey cells correspond to pathways that were not detected as significant or having a primary impact equal to 0.

CX1, suggesting the ability of cells to overcome incorrect spindle/kinetochore attachments[13]. APC/C inhibition is connected to the loss of its phosphorylation, required for its activation, and the inability to degrade its activator Cdh1 in G0 and G1 and cell cycle proteins prior to satisfaction of the cell cycle checkpoint. Genes involved in phosphorylation of the APC/C inhibitor Emi1, required for its degradation, are downregulated in BRCA and OV at CX1. Emi1 overexpression is a strong marker for CIN in solid tumors and it is frequently associated with aneuploidy[14].
E2F enabled inhibition of pre-replication complex formation is also negatively regulated in CX1 (BRCA, OV), as well as TP53 regulation of transcription of

genes involved in G1 cell cycle arrest (BRCA, UCEC, OV).
Pathways related to DNA replication initiation are also inactivated in CX1 (BRCA, UCEC), as well as CDC6 association with ORC origin complex and unwinding of DNA (BRCA, UCEC, OV).
Immune system response appears to be more active in CX3 with interleukin 36 pathway, TRAF6 mediated activation of interferon regulatory factor 7 (IRF7) (BRCA, UCEC) and cross presentation of soluble exogenous antigens (BRCA, UCEC, CESC) significantly altered. Release of apoptotic factors from the mitochondria is also present in CX3 (BRCA, OV).
The metabolism of high CX3 patients appears to be predominantly based on oxidative phosphorylation, as demonstrated by the upregulation of genes involved in respiratory electron transport and complex I biogenesis in CX3 for BRCA, UCEC and OV.
Interestingly, nuclear pore complex (NPC) disassembly and several pathways involved in protein import and export from nucleus, including TPR, are active in CX3 (BRCA, OV, CESC). mRNA trafficking between nucleus and cytoplasm is often aberrant in cancer, and this may result from chromosomal translocations affecting several nucleoporins, e.g. Nup98, TPR and RANBP2[15]. Pathways involved in RNA processing are also more active in CX3, e.g., mRNA decay and U12 dependent splicing of minor introns (BRCA, UCEC, OV), responsible for the correct splicing of several cancer related genes, including PTEN[16].
Finally, deficiencies of glycosyltrasferases are frequently detected as significant across analyses.

3.1.2 Genes

The same type of comparison is performed at the level of both primary and secondary genes detected across analyses at expression level. The heatmap in Figure 7 shows the top 10 primary genes for relevance and absolute value of logFC detected for each analysis.

Genes coding for nucleoporins, proteasome components, RNA polymerase II (L, E, G, H), and the oncogenes SRC and GRB2, tend to be upregulated in CX3 across tumors, except UCS.
DNA repair genes are also dysregulated, including NEIL3, involved in base-excision repair, overexpressed in CX3 (BRCA, OV, UCS), and genes participating in DNA double strand break repair (RAD50 and NBN). Genes involved in DNA damage sensing (TP53, ATM) are upregulated in CX1 across all tumors (TP53 was not detected as dysregulated in UCS). Genes involved in autophagy activation (BECN1, UVRAG), are downregulated in BRCA, OV, UCS in CX3. PIK3R1, believed to function as a tumor suppressor[17], is upregulated in CX1 across all tumors, except CESC.

Lysyl oxidase, LOX, and LOX-like genes, involved in extracellular matrix formation and associated with poor prognosis when overexpressed in gastric cancer[18], are also dysregulated.
The most divergent tumor is UCS that is not sharing many genes with other tumor types. Specific genes that were detected as primary in CX1 analysis are the peroxidase PXDN (downregulated), the proteoglycan GPC5 (downregu-

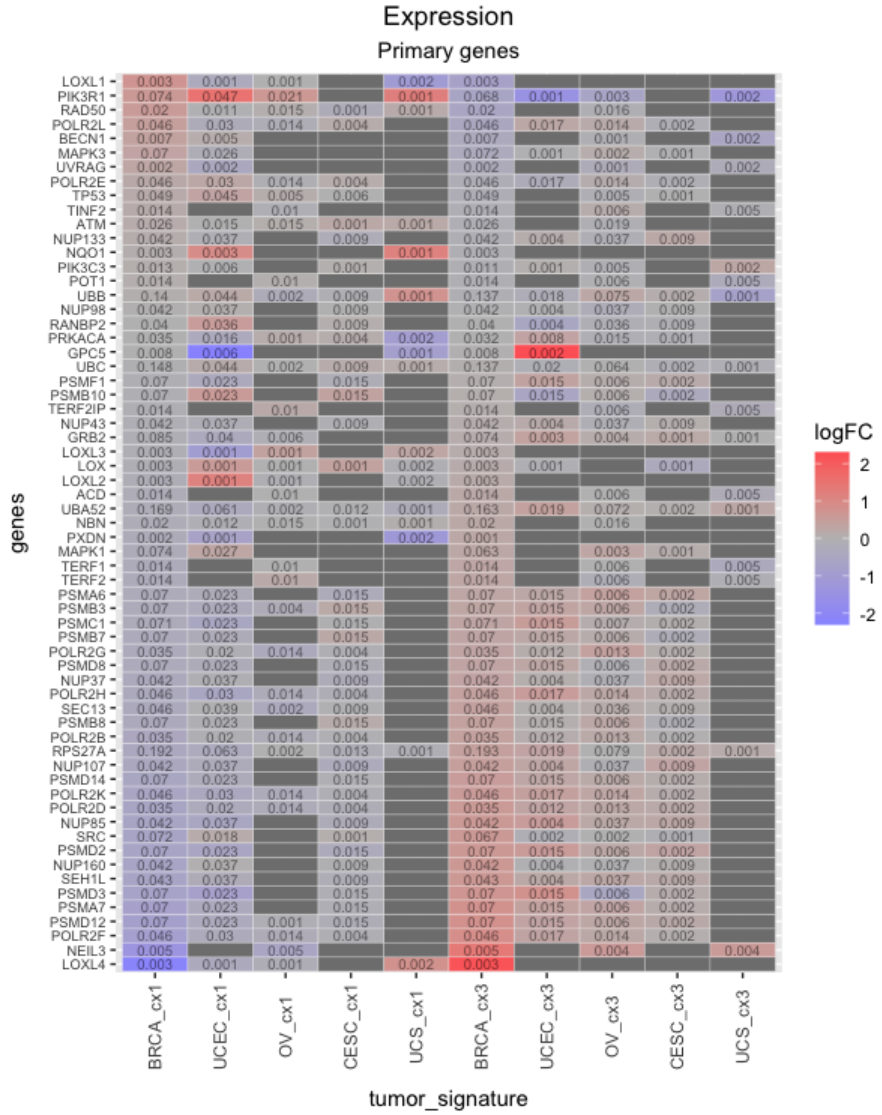


Figure 7: Heatmap showing the top 10 primary genes from Reactome pathways for relevance and logFC detected across analyses at expression level. The color of each cell corresponds to the logFC for the gene, while the number inside cells is the relevance. Grey cells correspond to genes that were not detected as primary in any pathway (relevance = 0).

lated) and the oxidoreductase NQO1 (upregulated). These genes were found also in UCEC with a similar behavior. In CX3 analysis genes with the highest relevance are involved in telomere function (TERF1/2, TINF2, POT1, ACD).

KEGG results (Appendix Figure 2) revealed tumor suppressors CDKN1A and RB1, apoptotic protein CASP9 are upregulated in CX1, while members of the E2F family of transcription factors are upregulated in CX3. Among the top 10 genes for relevance detected across analyses, 7 genes were found in both KEGG and Reactome analyses (NQO1, PIK3R1, TP53, MAPK1, MAPK3, SRC, PRKACA).

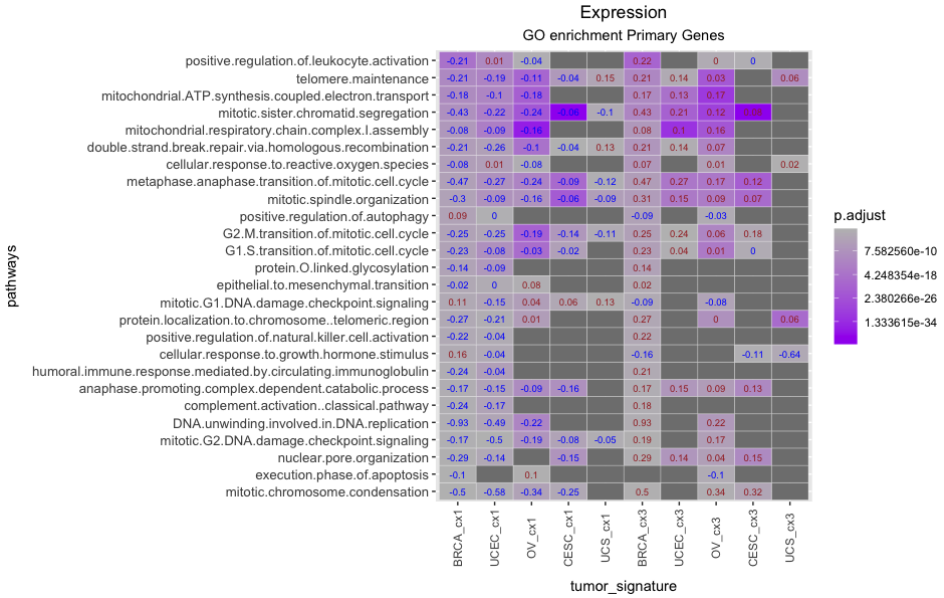
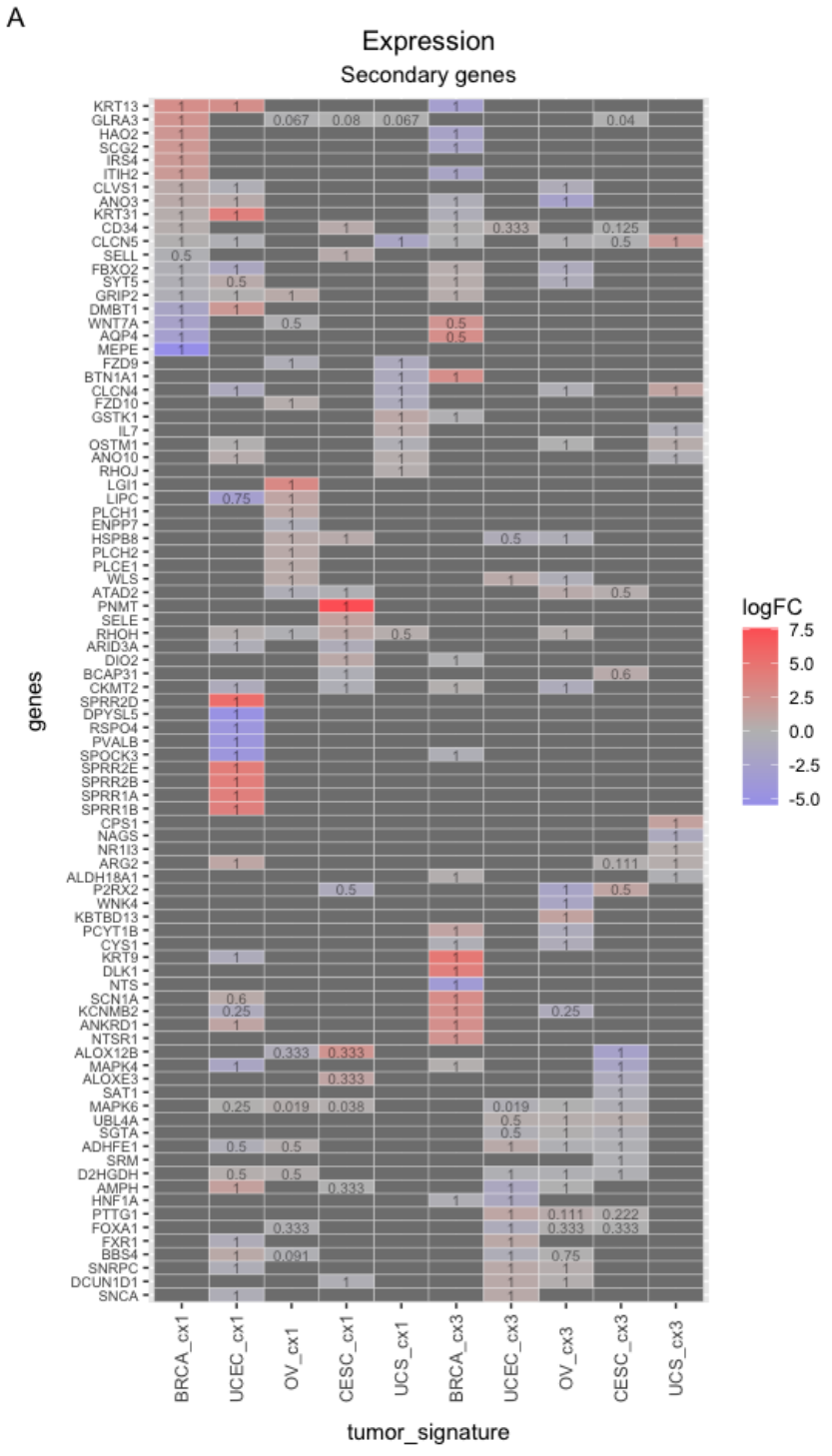


Figure 8: Heatmap showing a selection of significant GO terms enriched on primary genes detected across expression analyses. Each cell is colored according to the value of adjusted p-value. Inside each cell the median logFC of primary genes belonging to the gene set is depicted, colored in red or blue when it is positive or negative respectively. Dark grey cells represent non-significant GO terms.

Starting from the set of primary genes detected by *SourceSet*, an enrichment analysis on Gene Ontology is performed, focusing on biological processes where dysregulated genes are over-represented. While *SourceSet* detects affected pathways considering the effect of altered genes on downstream ones, in a graph structure, the enrichment analysis on Gene Ontology, instead, estimates the significance of a biological process based on how many genes from a given list of dysregulated genes belong to that process, without considering their level of dysregulation or their reciprocal interactions.

A large number of significant results were obtained from the enrichment analysis. The heatmap in Figure 8 shows some of the most interesting GO terms among the significantly enriched ones across analyses. They were selected based on interesting genes and pathways that were observed among *SourceSet* results. CX1 analyses show negative logFCs for several processes involved in chromosome segregation, including the metaphase/anaphase transition and mitotic spindle organization. DNA damage checkpoint at G2 is frequently inactivated in CX1 analyses.
Response to growth hormone is inactivated in CX3 (BRCA, CESC, UCS). DNA replication appears to be active in CX3 compared to signature 1. Respiratory electron transport is also active, as well as nuclear pore complex organization. Unexpectedly, homologous recombination is not inactivated in CX3; indeed, considering for example ovarian cancer results, several genes involved in homology-directed DNA repair are upregulated in CX3, with few exceptions, including BRCA1 (logFC = -0.26) and BARD1 (logFC = -0.11). Moreover, Reactome pathways relative to defective homologous recombination due to loss of BRCA1 and PALB2 functions are significant in CX3 analysis (primary.impact = 0.78).

Looking at the heatmap in Figure 9a, showing the top 10 secondary genes detected for each analysis, results appear to be more variable across tumors, compared to primary genes. Among the genes that were found dysregulated, there are immune response related genes (SELE, IL7, BTN1A1), several genes coding for ion channels (ANO3, CLCN4/5, P2RX2, SCN1A, KCNMB2), fatty acid metabolism (HAO2, ALOXE3, ALOX12B), transcription factors controlling cell proliferation and differentiation (FOXA1, ANKRD1, ARID3A), genes involved in keratinization (SPRR2D, KRT13) and ECM structure (ITIH2, SPOCK3), and, finally, PTTG1, that prevents separation of sister chromatid and is degraded by APC/C. Performing a GO enrichment analysis on secondary genes, it is possible to have a general overview of their main functions. Specific modules of enriched GO terms can be identified in Figure 9b (showing the top 10 most significant biological processes detected for each analysis), including calcium and cAMP signaling, ion transmembrane transport, regulation of tyrosine kinase activity, glycerophospholipid biosynthetic process, nucleotide metabolic process and RNA processing.



(a) Heatmap showing the top 10 secondary genes for total impact and absolute value of logFC detected across expression analyses. Cells are colored according to the logFC of the gene, inside each cell the total impact is depicted.

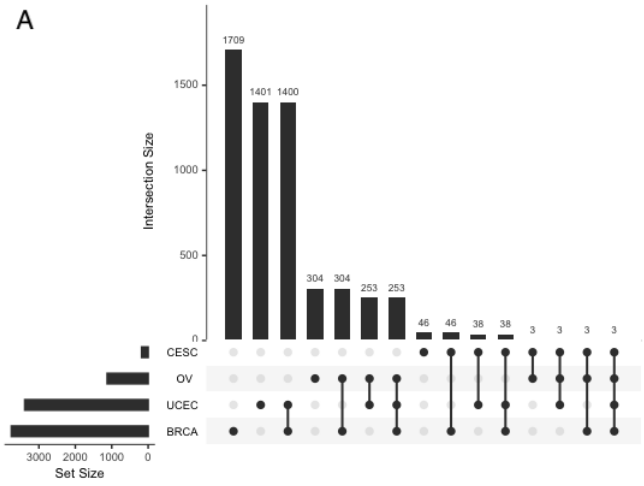
(b) Heatmap showing the top 10 GO terms enriched on secondary genes. Cells are colored according to the adjusted p-value.

Figure 9: Heatmaps relative to secondary genes.

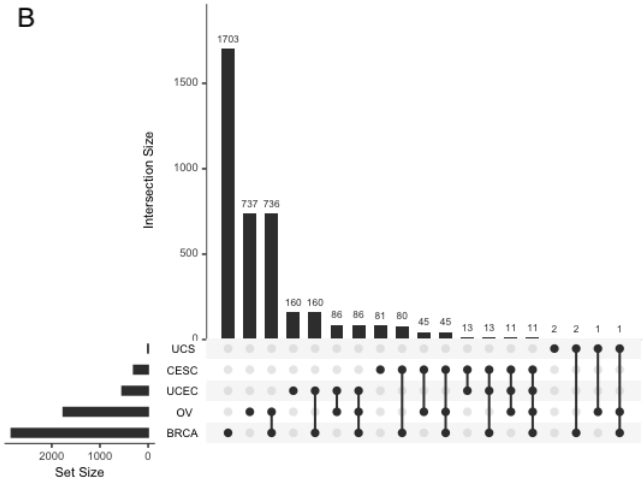
3.1.3 Comparison of significant pathways detected across tumors

While the previous heatmaps were relative only to the top 10 pathways detected for each analysis, it is also worth looking at the global results. Upset plots in Figures 10a, 10b show the intersections of significant Reactome pathways across each combination of tumors. BRCA, UCEC and OV are sharing the greatest amount of pathways in both CX1 and CX3 analyses (253, 86). CESC and UCS appear to be the most divergent tumors, with very few or zero significant pathways detected. Both of them are sharing more pathways with UCEC in CX1 analyses and more pathways with OV in CX3 analyses.

However, these findings may be related to the unbalancement on the size of the two groups of patients that have been compared, resulting in a decreased efficacy of the software on estimating the source set, as well as to the different amount of available samples across different tumors.



(a) Results relative to analyses on CX1.



(b) Results relative to analyses on CX3.

Figure 10: Intersection sizes of Reactome pathways detected as significant across each combination of tumors at expression level.

3.2 Methylation

3.2.1 Pathways

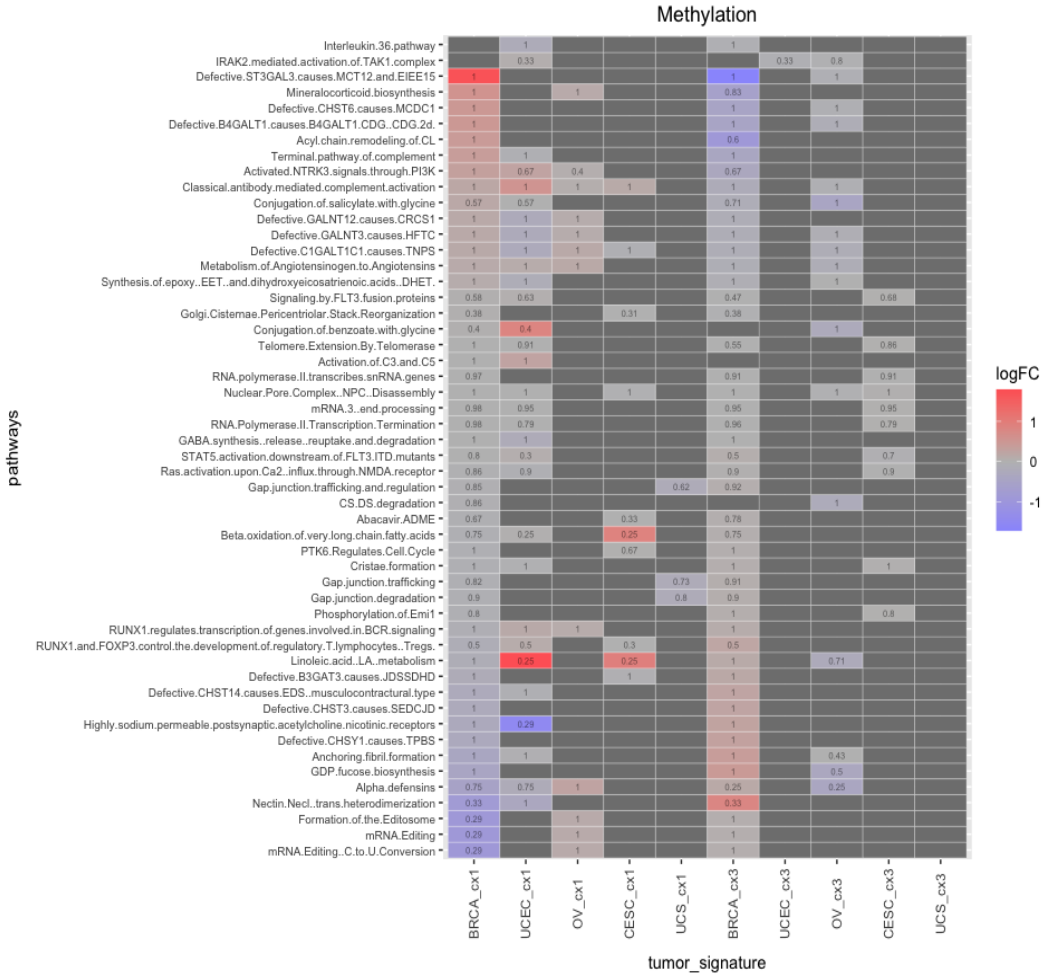


Figure 11: Heatmap showing the top 10 Reactome pathways for primary impact and absolute value of logFC detected across analyses at methylation level. The color of cells corresponds to the median logFC for the pathway, while the number inside cells is the primary impact. Grey cells correspond to pathways that were not detected as significant or having a primary impact equal to 0.

At the level of methylation (Figure 11), the classical antibody mediated activation of complement shows positive logFCs across cancer types in CX1. Several pathways involved in immune system response are dysregulated besides those relative to complement activation, including the alpha defensins (BRCA, UCEC and OV), the action of RUNX1 on the transcription of genes involved in BCR signaling and on the development of regulatory T lymphocytes (BRCA, UCEC, OV, CESC), the activation of TAK1 complex by IRAK2

(UCEC, OV).
mRNA editing, especially the C to U conversion, has a high primary impact in OV CX1. mRNA editing involving the dysregulation of APOBEC enzymes is frequently observed in cancer; these enzymes are also able to edit DNA introducing mutations and they are known to affect immune system response[19].
The peroxisomal metabolism of long chain fatty acids is altered in UCEC and CESC CX1, where genes belonging to linoleic acid metabolism and beta-oxidation of very long chain fatty acids tend to be hypermethylated. Cell adhesion is affected at the level of Nectins and Nectin-like dimerization in adherens junctions (negative logFCs in BRCA, UCEC CX1) and gap junctions (BRCA, UCS CX1).
Some pathways that were found affected starting from expression data were detected also by methylation analyses, such as PI3K activation of NTRK3 (positive logFCs in BRCA, UCEC and OV CX1), several pathways involved in glycosylation diseases, telomere extension by telomerase, nuclear pore complex disassembly and phosphorylation of Emi1.

3.2.2 Genes

Among genes that were found dysregulated both at the level of expression and methylation (Figure 12) were found proteasome components (UBC, UBA52, UBB), NPC components, telomere functioning genes, PI3KR1, NEIL3, GRB2. New genes that were not already detected in the previous analyses, include components of the transcriptional corepressor SMRT (TBL1X, HDAC3) and immune response (TRAF6, IRAK2, CR1), the DNA polymerase δ cofactor PCNA and microtubule associated EML4 (essential for mitotic spindle assembly and kinetochore attachment[20]).

Focusing on genes with the highest logFCs, interesting genes are the RNA binding protein KHSRP (CESC_cx1 = 1.15, BRCA_cx1 = 1.84), promoting metastasis and cell growth in non-small cell lung cancer[21], the hydroxylase involved in the formation of steroid hormones CYP21A2 (BRCA_cx1 = 3.62), the cell proliferation-promoting genes NEK6 (BRCA_cx1 = 3.28) and NEK7 (BRCA_cx1 = -4.07), the IL-1-receptor-associated kinase IRAK2 (UCEC_cx3 = -1.77), fibrinogen chains FGG (OV_cx3 = -4.80) and FGA (OV_cx3 = 2.50), growth factor FGF1 (OV_cx3 = 4.33), BDH2 (UCEC_cx1 = 3.03), involved in ketone bodies metabolism, genes involved in peroxisomal beta-oxidation of fatty acids ABCD1 (CESC_cx1 = 1.86) and ACOX1 (BRCA_cx1 = -1.04).
Comparing the top 10 genes detected across tumors between Reactome and KEGG (Appendix Figure 8), 9 genes overlap: GABBR1, GABBR2, GNB5, GNB3, MAPK1, MAPK3, PIK3CA, PIK3R1, ADCY1.

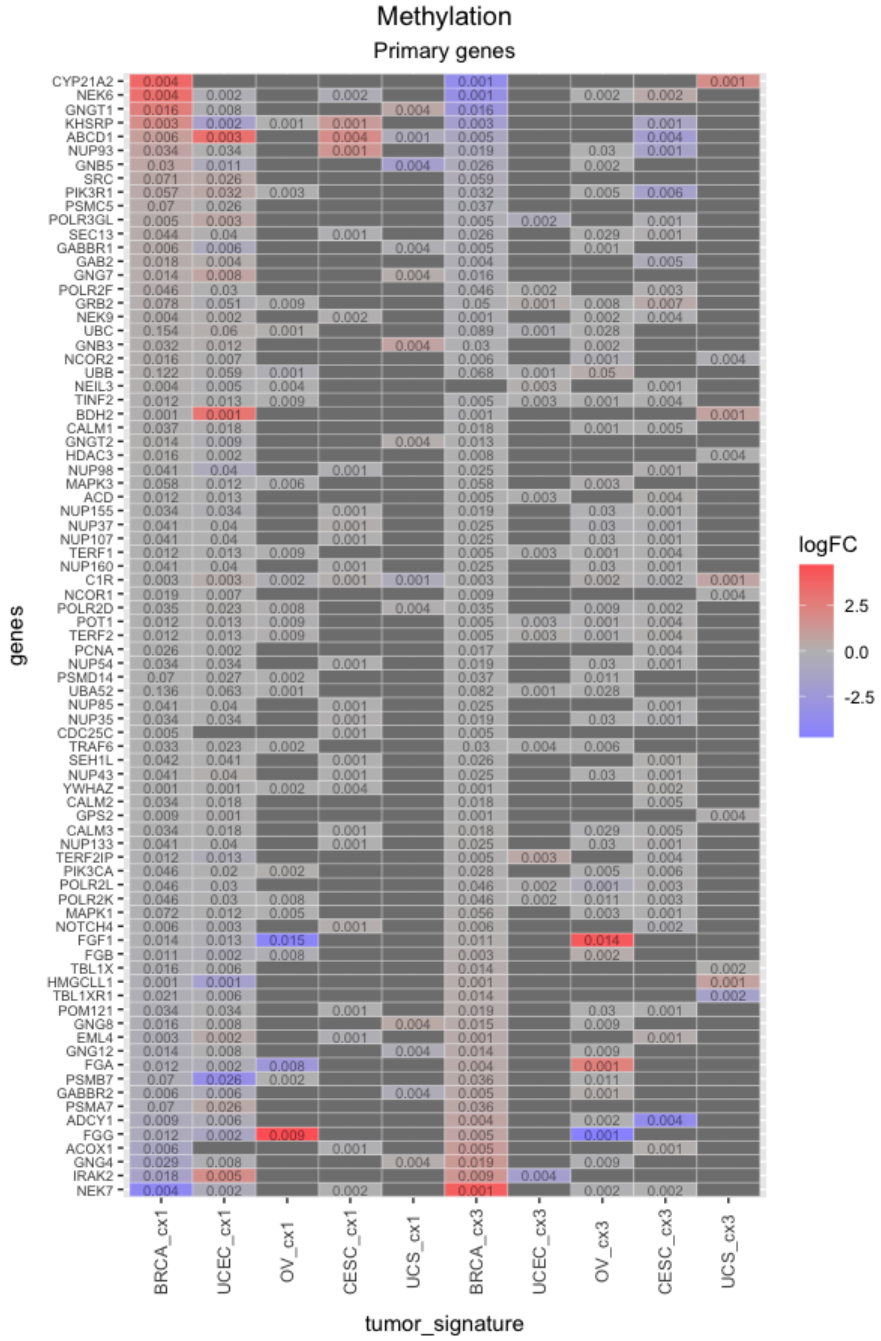


Figure 12: Heatmap showing the top 10 primary genes from Reactome pathways for relevance and logFC detected across analyses at methylation level. The color of each cell corresponds to the logFC for the gene, while the number inside cells is the relevance. Grey cells correspond to genes that were not detected as primary in any pathway (relevance = 0).

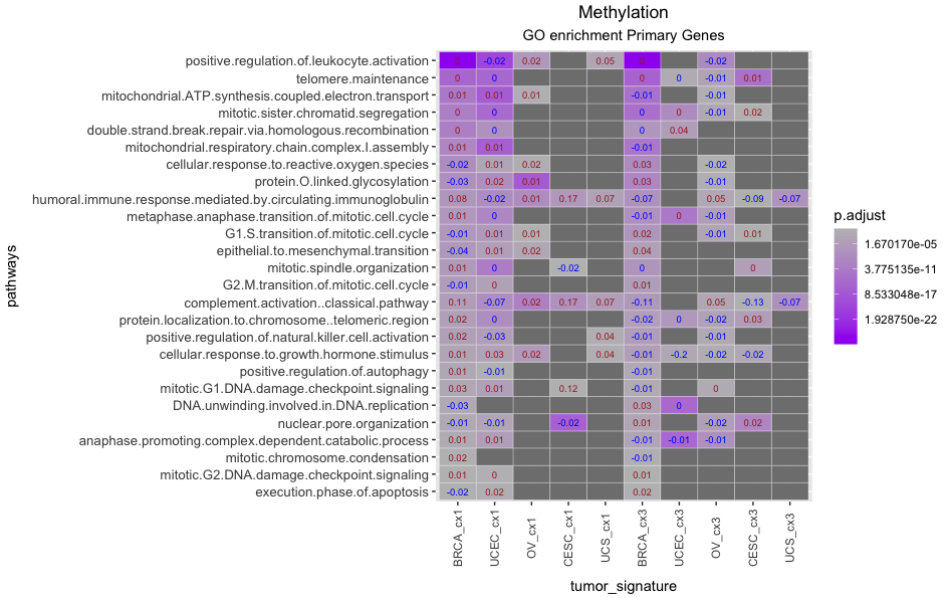
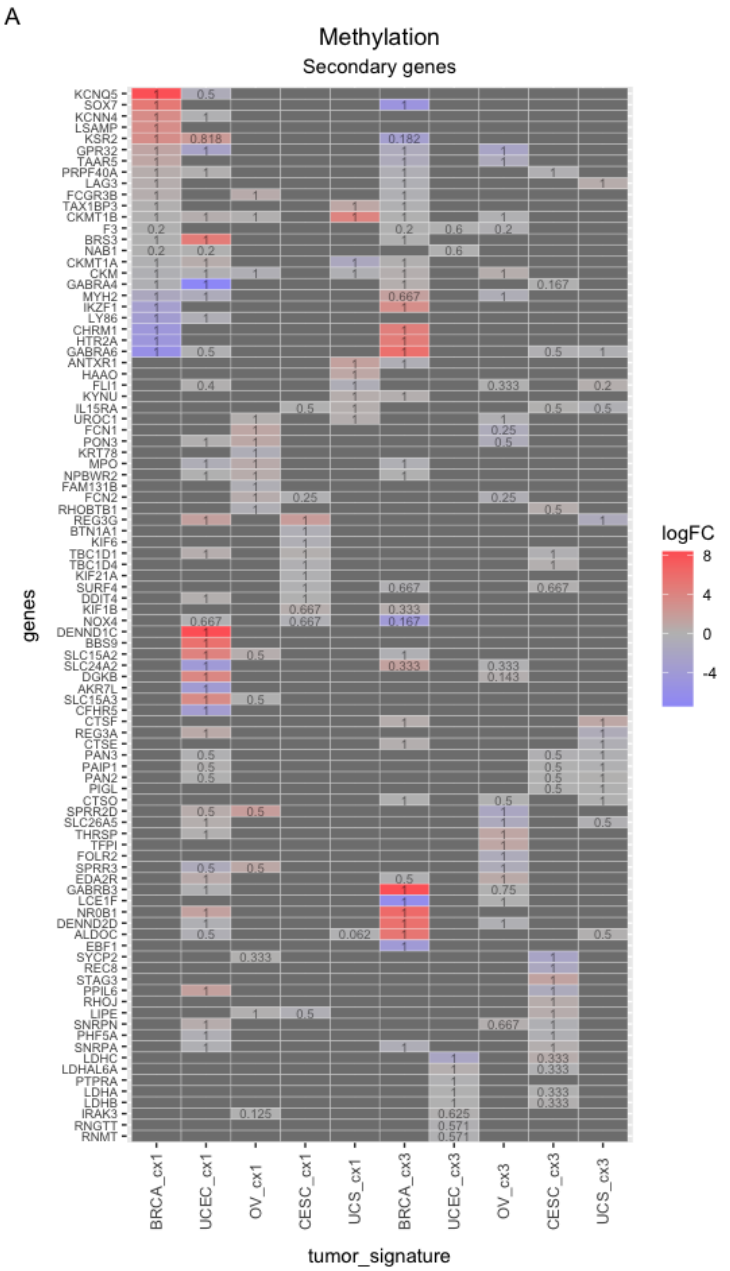


Figure 13: Heatmap showing a selection of significant GO terms enriched on primary genes detected across methylation analyses. Each cell is colored according to the value of adjusted p-value. Inside each cell the median logFC of primary genes belonging to the gene set is depicted, colored in red or blue when it is positive or negative respectively. Dark grey cells represent non-significant GO terms.

The heatmap (Figure 13) shows the results for the enrichment analysis on primary genes detected by *SourceSet* from methylation data for the same selection of GO biological processes that was considered for expression data. Many biological processes are still significant at the level of methylation; however values of logFC are close to zero for the majority of the pathways. Interestingly, complement activation and huomral immune response mediated by immunoglobulin have slightly positive logFC across tumors for signature 1, with the exception of UCEC.

Secondary genes detected across methylation analyses (Figures 14a, 14b) are less overlapping between tumor types compared to primary genes, confirming the higher variability of these genes that was previously observed at the level of expression. The most represented processes include ion channels (KCNN4, KCNQ5, GABRA4), immune system response (LAG3, FCGR3B, FCN1), meiotic chromosome organization (SYCP2, REC8, STAG3), tumor metabolism (CKMT1A/B, THRSP, LDHA, LDHB, LDHC), cell adhesion (RHOJ, ANTXR1, TAX1BP3).

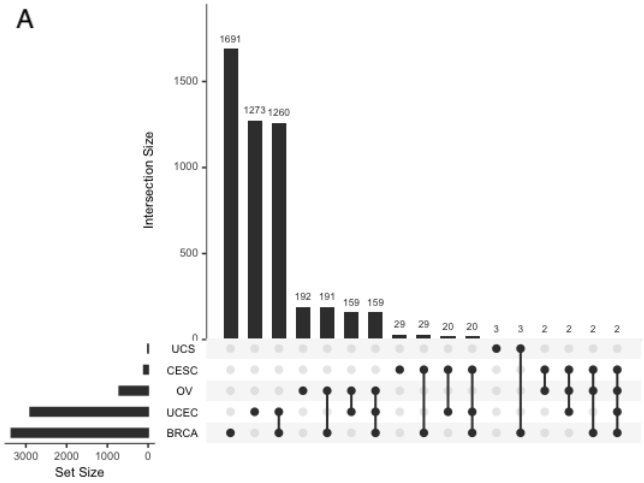


(a) Heatmap showing top 10 secondary genes for total impact and absolute value of logFC detected across methylation analyses. Cells are colored according to the logFC of the gene, inside each cell the total impact is depicted.

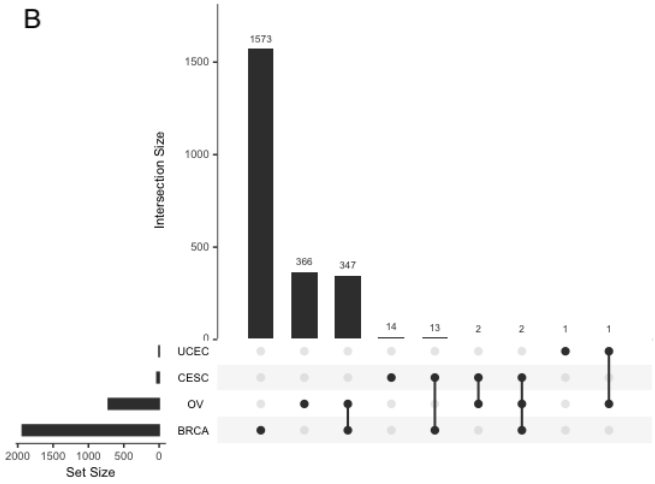
(b) Heatmap showing the top 10 GO terms enriched on secondary genes. Cells are colored according to the adjusted p-value.

Figure 14: Heatmaps relative to secondary genes.

3.2.3 Comparison of significant pathways detected across tumors



(a) Results relative to analyses on CX1.



(b) Results relative to analyses on CX3.

Figure 15: Intersection sizes of Reactome pathways detected as significant across each combination of tumors at methylation level.

At the level of methylation, the number of significant pathways detected by *SourceSet* (Figures 15a, 15b) is slightly lower compared to analyses on expression data. Again, for signature 1 BRCA and UCEC are the most similar (1260), and they also have many pathways in common with OV (159). The amount of significant pathways detected for CESC and UCS is limited or null, with also UCEC having only one significant pathway in CX3 analysis.

3.3 Expression VS Methylation: anticorrelated genes

Among genes that were detected from both expression and methylation data (Table 6), a proportion (45% - 50%) show anticorrelated expression and methylation profiles, with opposite signs of logFC; however, for many of these genes the differential level of expression or methylation was very small. A threshold of 0.1 is applied to select genes with a reasonably high absolute value of logFC for both omics.

Reactome (primary genes)

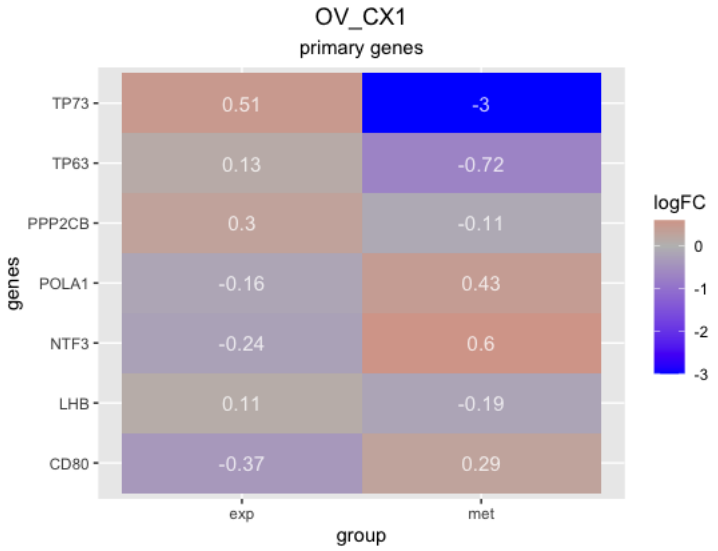
		Expression	Methylation	Expression ∩ Methylation	Anticorrelated
CX1	BRCA	7034	6522	5840	2692
	UCEC	4166	4058	2854	1346
	OV	978	811	194	87
	CESC	1114	108	33	16
	UCS	164	146	1	1
CX3	BRCA	6667	5469	4902	2234
	UCEC	1225	129	59	24
	OV	2286	1131	532	263
	CESC	578	614	122	56
	UCS	103	37	1	0

Table 6: Number of primary genes detected in expression and methylation analyses on Reactome, the amount of genes that were found in both omics and the number of anticorrelated genes which show opposite logFC signs between expression and methylation.

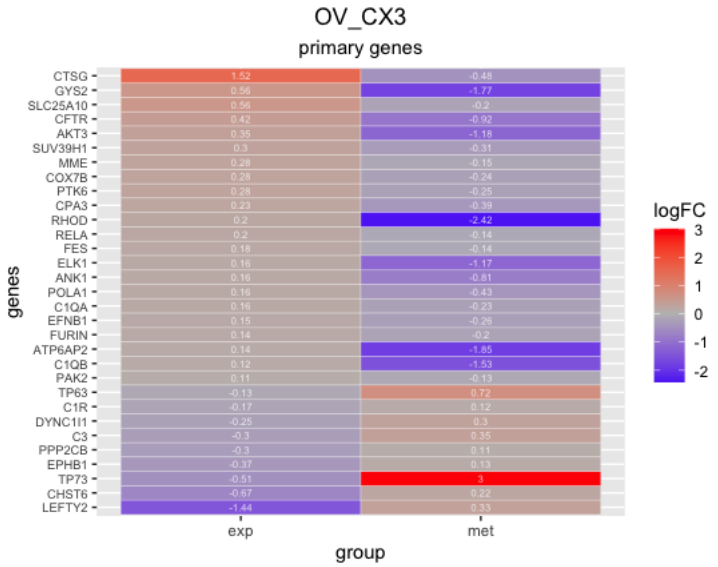
3.3.1 Anticorrelated genes in ovarian cancer

Considering the results relative to ovarian cancer, tumor suppressors TP73 and TP63 are hypomethylated genes with increased expression in patients with high CX1 levels (Figure 16a). NTF3, CD80 and POLA1 show opposite behavior, being hypermethylated and underexpressed in patients bearing high levels of signature 1.
In CX3 analysis (Figure 16b), among genes whose anticorrelated expression and methylation levels have been already associated with cancer there are CFTR[22], PTK6[23], RHOD[24], FES[25], ANK1[26], LEFTY2[27]. Complement system components frequently show opposite logFC (C1QA, C1QB, C3, C1R). The oncogene FURIN was also found in CESC analysis for signature 3, with opposite behavior.

Interestingly, the already mentioned IRAK2 and the oncogene ERBB2 were among the anticorrelated genes in UCEC. Additional plots relative to the other tumors are available in Appendix.



(a) Results relative to CX1.



(b) Results relative to CX3.

Figure 16: Heatmaps showing anticorrelated primary genes with logFC $>$ 0.1 or logFC $<$ -0.1 for both expression and methylation, relative to ovarian cancer analyses. Cells are colored according to the logFC, that is also indicated inside.

4 Discussion

The topological pathway analysis performed by *SourceSet* was able to identify and confirm the main biological mechanisms that are responsible for a specific pattern of chromosomal instability. The cause of the observed perturbation is captured on the source set of analyzed pathways, with coherent results between KEGG and Reactome analyses.
Primary genes, responsible for the perturbation associated with a specific signature, are often more conserved across tumor types, reflecting the presence of a specific aetiology for the signature. Primary genes tend to have smaller values of logFC compared to secondary genes and often the top primary genes for logFC do not correspond to the top genes for relevance. Indeed, even a small perturbation on these genes is able to induce massive dysregulations on downstream genes through network propagation. Moreover, secondary genes present higher variability, being involved in a large number of different processes, and they are not homogeneously distributed across analyses, reflecting the tissue-specificity of expression profiles.

Comparing *SourceSet* results for expression and methylation data, methylation results are generally more variable across different tumor types. logFC values are frequently close to zero, with few exceptions of very high values, indicating that the changes in methylation levels across analyses tend to be mild, with rare cases of strong modifications. Several genes and pathways were found affected in both omics, especially related to nuclear pore complex organization, telomere function, proteasome components, confirming the role of the methylation status of a gene in the regulation of its expression in cancer. However, while expression profiles are indeed affected by genomic instability, it is still unclear and worthy of further investigation whether the alteration of methylation profiles develops as a consequence of genomic instability or it is contributing to the onset of CIN.

Results obtained from both expression and methylation data confirm the closeness of BRCA, UCEC and OV molecular profiles, compared to CESC and UCS. Although the ability of the software in identifying the source set when the number of samples is limited can be compromised, these findings may also have a biological meaning. Indeed, CESC and UCS have important features that differentiate them from other tumor types: cervical cancer predominantly originates from viral infection by HPV and uterine carcinosarcoma is not a proper carcinoma like other tumors, since it contains an important mesenchymal component.

The two signatures of CIN are clearly generated by different dysregulated mechanisms that *SourceSet* was able to capture.

Signature 1 is characterized by defects during mitosis, especially at the spindle assembly checkpoint, when, even in the presence of an incorrect attachment between kinetochore and microtubules, the cell is able to divide without entering apoptosis. This is a process that involves the APC/C complex in the metaphase/anaphase transition. APC/C activated by Cdc20 is responsible for securin ubiquitination leading to its degradation and subsequent chromosome separation. The activation of the spindle assembly checkpoint recruits protein complexes that inhibit APC/C, while also decreasing the cytoplasmic pool of APC/C-Cdc20[13]. In patients bearing high levels of CX1, APC/C is not properly inhibited to avoid entering anaphase with incorrect chromosome segregation, resulting in the observed pattern of CIN, i.e., aneuploidy. The incorrect attachment of microtubules to the kinetochore may occur by altered deposition of CENPA nucleosomes at centromeres, essential for kinetochore attachment[28], and defects in the organization of centrosomes, that are responsible for spindle assembly through microtubule enucleation[29].

Patients with high signature 3 are instead characterized by long-sized single-copy changes, suggesting impaired homologous recombination. Unexpectedly, genes or pathways involved in this type of DNA repair mechanism are not observed among the dysregulated ones with high relevance or primary impact; moreover, they are often overexpressed in this signature. Indeed, the aetiology of CX3 also involves replication stress, represented by stalled replication forks, that, being unstable, may generate single or double strand breaks if not properly protected. Defects involving DNA replication are normally detected in S phase through ATM/ATR that will activate DNA repair pathways, preferentially homologous recombination, and, thus, genes responsible for this DNA repair pathway are upregulated in this signature. However, homologous recombination may be impaired due to loss-of-function mutations on essential components for this pathway, e.g., BRCA1/2, RAD51C, PALB2[30]. DNA lesions, thus, must be repaired with alternative pathways, such as non-homologous end joining and theta-mediated end joining, which are error-prone and contribute to genomic instability. Replication stress in CX3 appears to derive from unscheduled DNA replication, generated when the timing of origin activation is altered, e.g., due to expression of an oncogene; indeed, DNA replication initiation mediated by CDC6 is upregulated in CX3. CDC6 overexpression is frequently responsible for replication stress and is associated with double strand breaks and genomic instability in cancer[31]. As a result, portions of DNA may be replicated more than once and the presence of multiple replication forks increases the risk of collisions and stalling, which may also occur in case of DNA damage. The presence of high levels of DNA replication suggests increased proliferation rates for tumoral cells bearing this signature, with higher demand of energy and dNTPs[32].
Finally, the two signatures also differentiate for immune response and energy

metabolism: signature 3 shows increased energy production by oxidative phosphorylation and higher immune response, whose related genes are frequently regulated by methylation levels.

In future studies, it would be interesting to extend the analysis by *SourceSet* to other signatures of chromosomal instability that are active in gynecological tumors. For example, signature 2 and 5, also associated with impaired homologous recombination, can be studied to determine how they differ from each other and from signature 3. Signature 4 and signature 10 are also slightly active, caused by deregulation of PI3K-AKT pathway and impaired NHEJ respectively. The same analysis can be extended to other tumors in a pan-cancer study, in order to better understand the role of these signatures in cancer, but also to further dissect the molecular characteristics of gynecological cancers with respect to non-gynecological ones. Finally, other omics can be analyzed, e.g., proteomics and CNV, to better characterize tumors and the effects of different signatures on them.

References

1. Siegel, R. L., Miller, K. D., Wagle, N. S. & Jemal, A. Cancer statistics, 2023. *CA: A Cancer Journal for Clinicians* **73,** 17–48. ISSN: 1542-4863. https://onlinelibrary.wiley.com/doi/abs/10.3322/caac.21763 (2023).
2. Bell, D. *et al.* Integrated genomic analyses of ovarian carcinoma. *Nature* **474.** Number: 7353 Publisher: Nature Publishing Group, 609–615. ISSN: 1476-4687. https://www.nature.com/articles/nature10166 (June 2011).
3. Levine, D. A. Integrated genomic characterization of endometrial carcinoma. *Nature* **497.** Number: 7447 Publisher: Nature Publishing Group, 67–73. ISSN: 1476-4687. https://www.nature.com/articles/nature12113 (May 2013).
4. Koboldt, D. C. *et al.* Comprehensive molecular portraits of human breast tumours. *Nature* **490.** Number: 7418 Publisher: Nature Publishing Group, 61–70. ISSN: 1476-4687. https://www.nature.com/articles/nature11412 (Oct. 2012).
5. Burk, R. D. *et al.* Integrated genomic and molecular characterization of cervical cancer. *Nature* **543.** Number: 7645 Publisher: Nature Publishing Group, 378–384. ISSN: 1476-4687. https://www.nature.com/articles/nature21386 (Mar. 2017).
6. Cherniack, A. D. *et al.* Integrated Molecular Characterization of Uterine Carcinosarcoma. *Cancer Cell* **31.** Publisher: Elsevier, 411–423. ISSN: 1535-6108, 1878-3686. https://www.cell.com/cancer-cell/abstract/S1535-6108(17)30053-3 (Mar. 13, 2017).
7. Sansregret, L., Vanhaesebroeck, B. & Swanton, C. Determinants and clinical implications of chromosomal instability in cancer. *Nature Reviews Clinical Oncology* **15.** Number: 3 Publisher: Nature Publishing Group, 139–150. ISSN: 1759-4782. https://www.nature.com/articles/nrclinonc.2017.198 (Mar. 2018).
8. Drews, R. M. *et al.* A pan-cancer compendium of chromosomal instability. *Nature* **606,** 976–983. ISSN: 1476-4687 (June 2022).
9. Nishiyama, A. & Nakanishi, M. Navigating the DNA methylation landscape of cancer. *Trends in Genetics* **37.** Publisher: Elsevier, 1012–1027. ISSN: 0168-9525. https://www.cell.com/trends/genetics/abstract/S0168-9525(21)00130-X (Nov. 1, 2021).

10. García-Campos, M. A., Espinal-Enríquez, J. & Hernández-Lemus, E. Pathway Analysis: State of the Art. *Frontiers in Physiology* **6,** 383. ISSN: 1664-042X. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4681784/ (Dec. 17, 2015).

11. Salviato, E., Djordjilović, V., Chiogna, M. & Romualdi, C. SourceSet: A graphical model approach to identify primary genes in perturbed biological pathways. *PLoS computational biology* **15,** e1007357. ISSN: 1553-7358 (Oct. 2019).

12. Berger, A. C. *et al.* A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers. *Cancer Cell* **33,** 690–705.e9. ISSN: 1878-3686 (Apr. 9, 2018).

13. Bharadwaj, R. & Yu, H. The spindle checkpoint, aneuploidy, and cancer. *Oncogene* **23.** Number: 11 Publisher: Nature Publishing Group, 2016–2027. ISSN: 1476-5594. https://www.nature.com/articles/1207374 (Mar. 2004).

14. Vaidyanathan, S. *et al.* In vivo overexpression of Emi1 promotes chromosome instability and tumorigenesis. *Oncogene* **35.** Number: 41 Publisher: Nature Publishing Group, 5446–5455. ISSN: 1476-5594. https://www.nature.com/articles/onc201694 (Oct. 2016).

15. Borden, K. L. B. The Nuclear Pore Complex and mRNA Export in Cancer. *Cancers* **13,** 42. ISSN: 2072-6694. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7796397/ (Dec. 25, 2020).

16. Nishimura, K., Yamazaki, H., Zang, W. & Inoue, D. Dysregulated minor intron splicing in cancer. *Cancer Science* **113,** 2934–2942. ISSN: 1347-9032. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9459249/ (Sept. 2022).

17. Liu, Y. *et al.* Pan-cancer analysis on the role of PIK3R1 and PIK3R2 in human tumors. *Scientific Reports* **12,** 5924. ISSN: 2045-2322 (Apr. 8, 2022).

18. Zhu, J. *et al.* Expression of LOX Suggests Poor Prognosis in Gastric Cancer. *Frontiers in Medicine* **8,** 718986. ISSN: 2296-858X. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8476844/ (Sept. 14, 2021).

19. Kurkowiak, M. *et al.* The effects of RNA editing in cancer tissue at different stages in carcinogenesis. *RNA Biology* **18,** 1524–1539. ISSN: 1547-6286. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8582992/.

20. Chen, D. *et al.* EML4 promotes the loading of NUDC to the spindle for mitotic progression. *Cell Cycle (Georgetown, Tex.)* **14,** 1529–1539. ISSN: 1551-4005 (2015).

21. Yan, M. *et al.* RNA-binding protein KHSRP promotes tumor growth and metastasis in non-small cell lung cancer. *Journal of experimental & clinical cancer research: CR* **38,** 478. ISSN: 1756-9966 (Nov. 27, 2019).

22. Wang, Y. *et al.* DNA Methylation-Mediated Low Expression of CFTR Stimulates the Progression of Lung Adenocarcinoma. *Biochemical Genetics* **60,** 807–821. ISSN: 1573-4927 (Apr. 2022).

23. Hsieh, Y.-P. *et al.* Epigenetic Deregulation of Protein Tyrosine Kinase 6 Promotes Carcinogenesis of Oral Squamous Cell Carcinoma. *International Journal of Molecular Sciences* **23,** 4495. ISSN: 1422-0067 (Apr. 19, 2022).

24. Duong, C. V. *et al.* Quantitative, genome-wide analysis of the DNA methylome in sporadic pituitary adenomas. *Endocrine-Related Cancer* **19,** 805–816. ISSN: 1479-6821 (Dec. 2012).

25. Shaffer, J. M. & Smithgall, T. E. Promoter methylation blocks FES protein-tyrosine kinase gene expression in colorectal cancer. *Genes, Chromosomes & Cancer* **48,** 272–284. ISSN: 1098-2264 (Mar. 2009).

26. Omura, N. *et al.* Overexpression of ankyrin1 promotes pancreatic cancer cell growth. *Oncotarget* **7,** 34977–34987. ISSN: 1949-2553 (June 7, 2016).

27. Gao, X., Cai, Y. & An, R. miR-215 promotes epithelial to mesenchymal transition and proliferation by regulating LEFTY2 in endometrial cancer. *International Journal of Molecular Medicine* **42,** 1229–1236. ISSN: 1107-3756. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6089757/ (Sept. 2018).

28. Yatskevich, S. *et al.* Structure of the human inner kinetochore bound to a centromeric CENP-A nucleosome. *Science (New York, N.Y.)* **376,** 844–852. ISSN: 1095-9203 (May 20, 2022).

29. Vasquez-Limeta, A. & Loncarek, J. Human centrosome organization and function in interphase and mitosis. *Seminars in cell & developmental biology* **117,** 30–41. ISSN: 1084-9521. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8465925/ (Sept. 2021).

30. Hoppe, M. M., Sundar, R., Tan, D. S. P. & Jeyasekharan, A. D. Biomarkers for Homologous Recombination Deficiency in Cancer. *JNCI: Journal of the National Cancer Institute* **110,** 704–713. ISSN: 0027-8874. https://doi.org/10.1093/jnci/djy085 (July 1, 2018).

31. Liontos, M. *et al.* Deregulated Overexpression of hCdt1 and hCdc6 Promotes Malignant Behavior. *Cancer Research* **67,** 10899–10909. ISSN: 0008-5472. https://doi.org/10.1158/0008-5472.CAN-07-2837 (Nov. 15, 2007).

32. Gaillard, H., García-Muse, T. & Aguilera, A. Replication stress and cancer. *Nature Reviews Cancer* **15.** Number: 5 Publisher: Nature Publishing Group, 276–289. ISSN: 1474-1768. https://www.nature.com/articles/nrc3916 (May 2015).

A Appendix

A.1 KEGG results

A.1.1 Expression

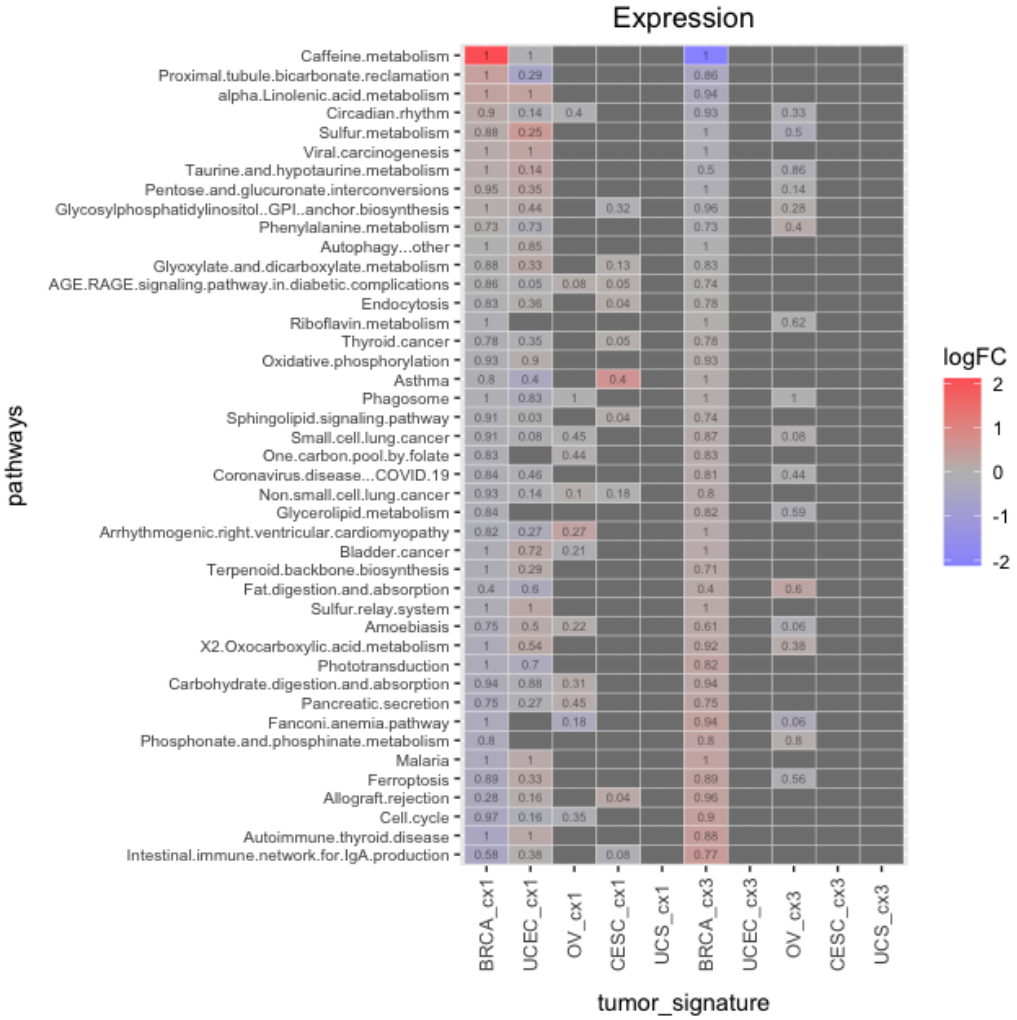


Figure 1: Heatmap showing the top 10 KEGG pathways for primary impact and absolute value of logFC detected across analyses at expression level. The color of cells corresponds to the median logFC for the pathway, while the number inside cells is the primary impact. Grey cells correspond to pathways that were not detected as significant or having a primary impact equal to 0.

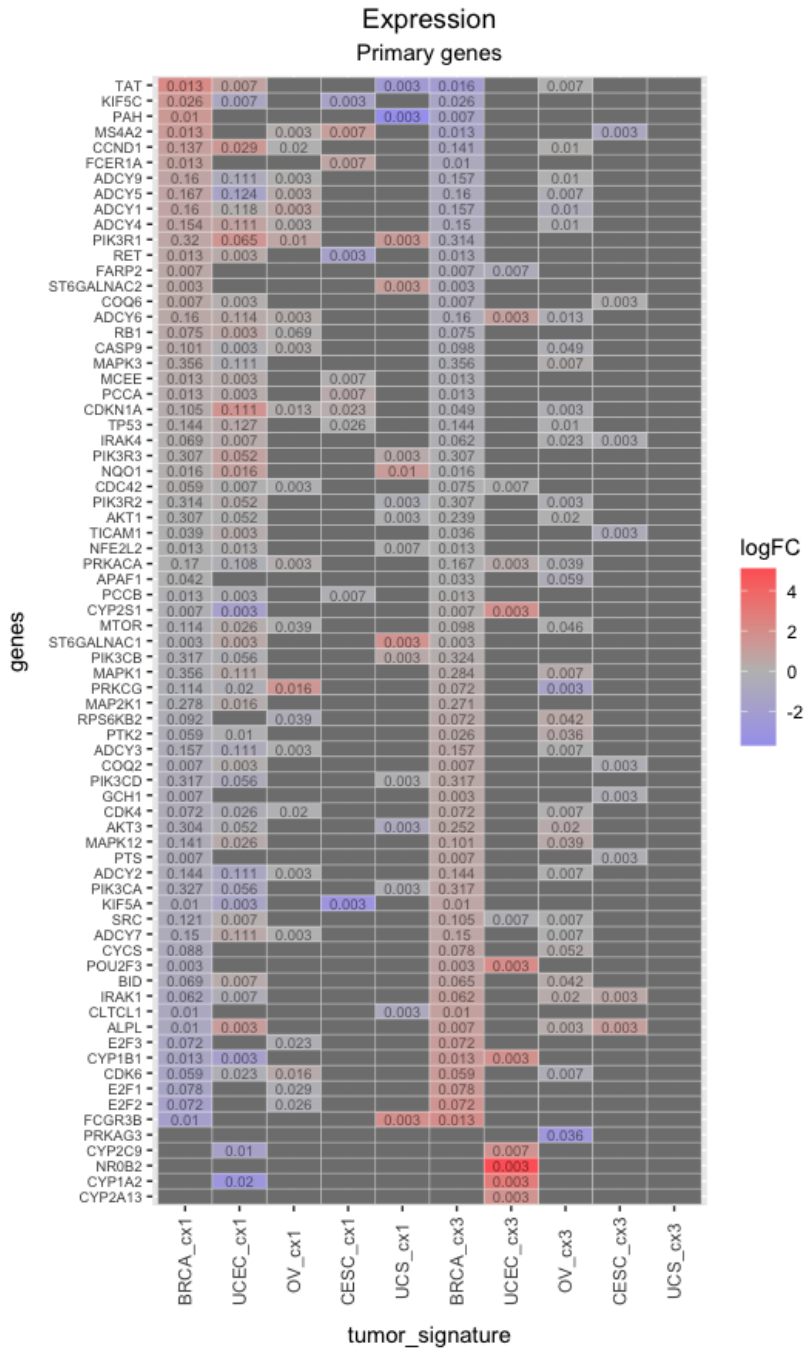


Figure 2: Heatmap showing the top 10 primary genes from KEGG pathways for relevance and logFC detected across analyses at expression level. The color of each cell corresponds to the logFC for the gene, while the number inside cells is the relevance. Grey cells correspond to genes that were not detected as primary in any pathway (relevance = 0).

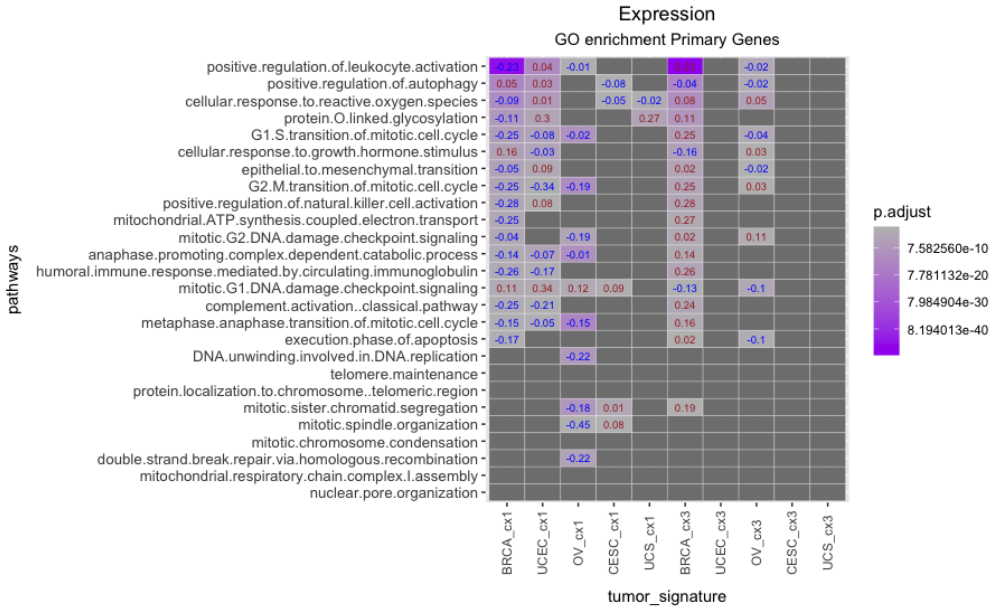


Figure 3: Heatmap showing a selection of significant GO terms enriched on primary genes detected across expression analyses. Each cell is colored according to the value of adjusted p-value. Inside each cell the median logFC of primary genes belonging to the gene set is depicted, colored in red or blue when it is positive or negative respectively. Dark grey cells represent non-significant GO terms.

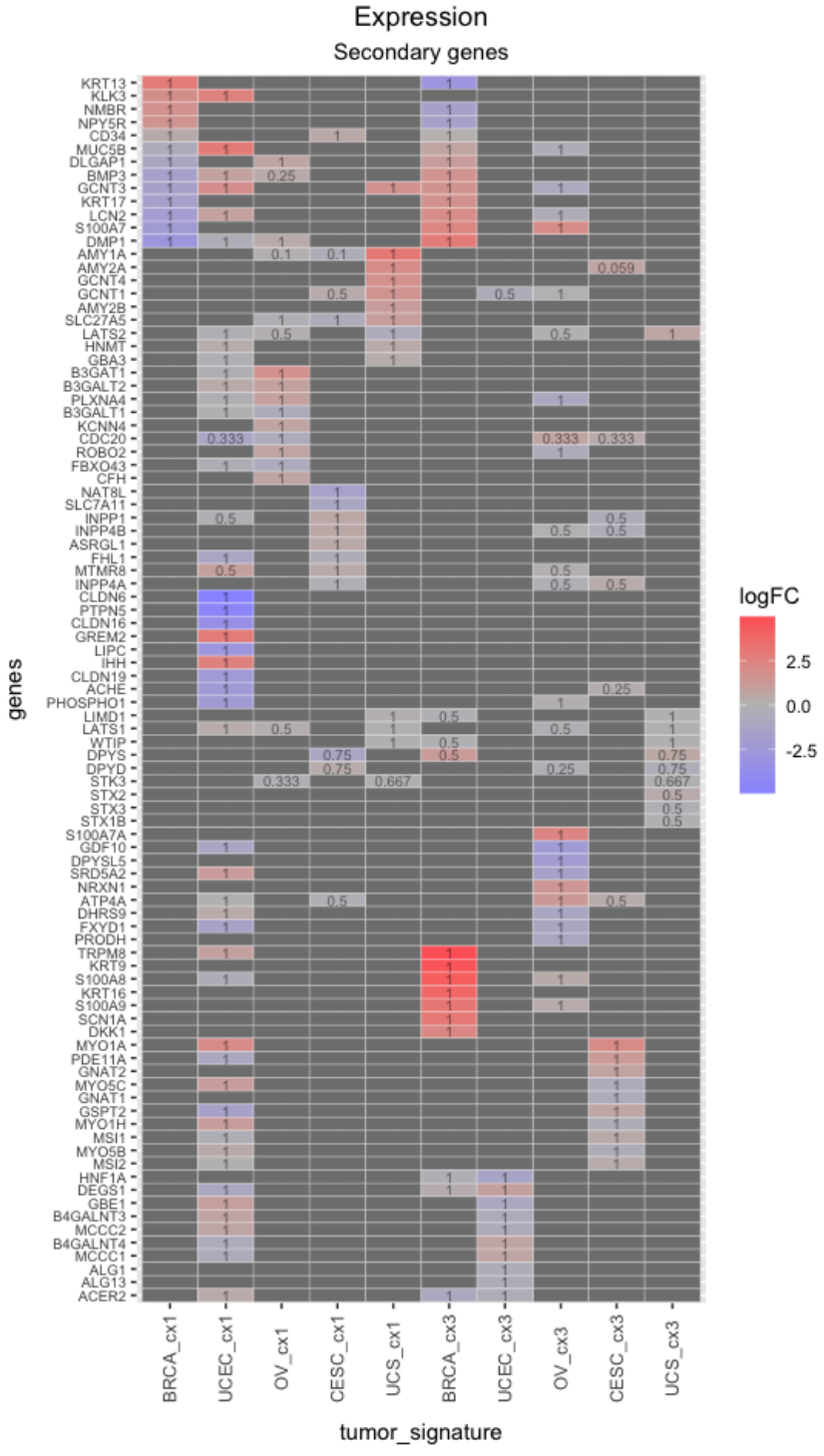


Figure 4: Heatmap showing the top 10 secondary genes for total impact and absolute value of logFC detected across expression analyses. Cells are colored according to the logFC of the gene, inside each cell the total impact is depicted.

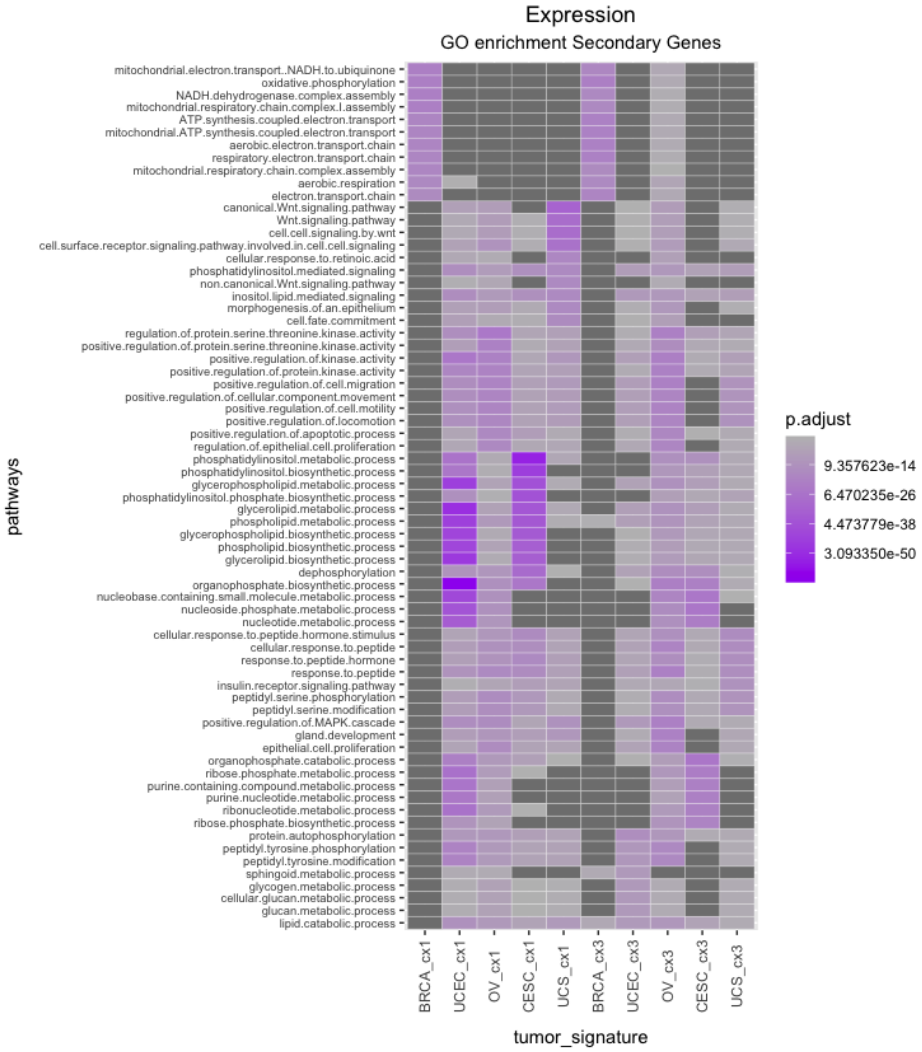
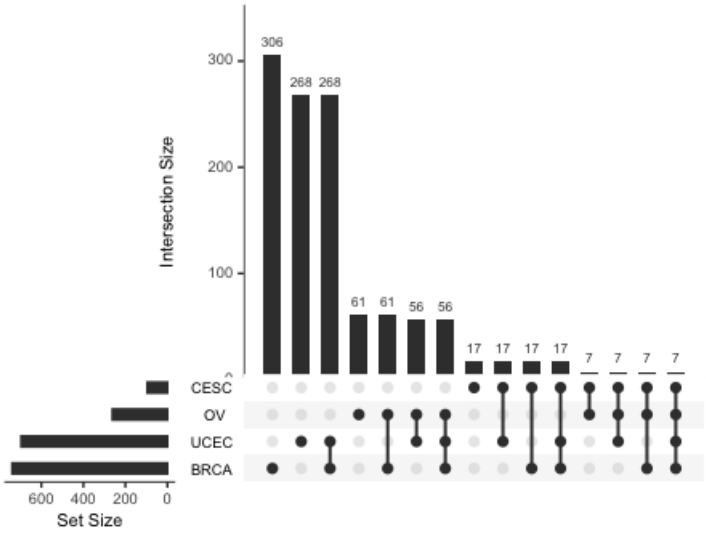
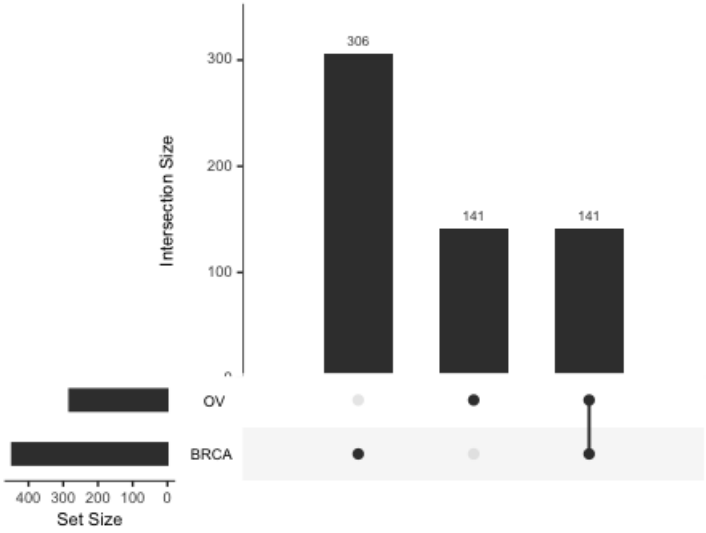


Figure 5: Heatmap showing the top 10 GO terms enriched on secondary genes. Cells are colored according to the adjusted p-value.



(a) Intersection sizes of KEGG pathways detected as significant across each combination of tumors at expression level for analyses on CX1.



(b) Intersection sizes of KEGG pathways detected as significant across each combination of tumors at expression level for analyses on CX3.

Figure 6

A.1.2 Methylation

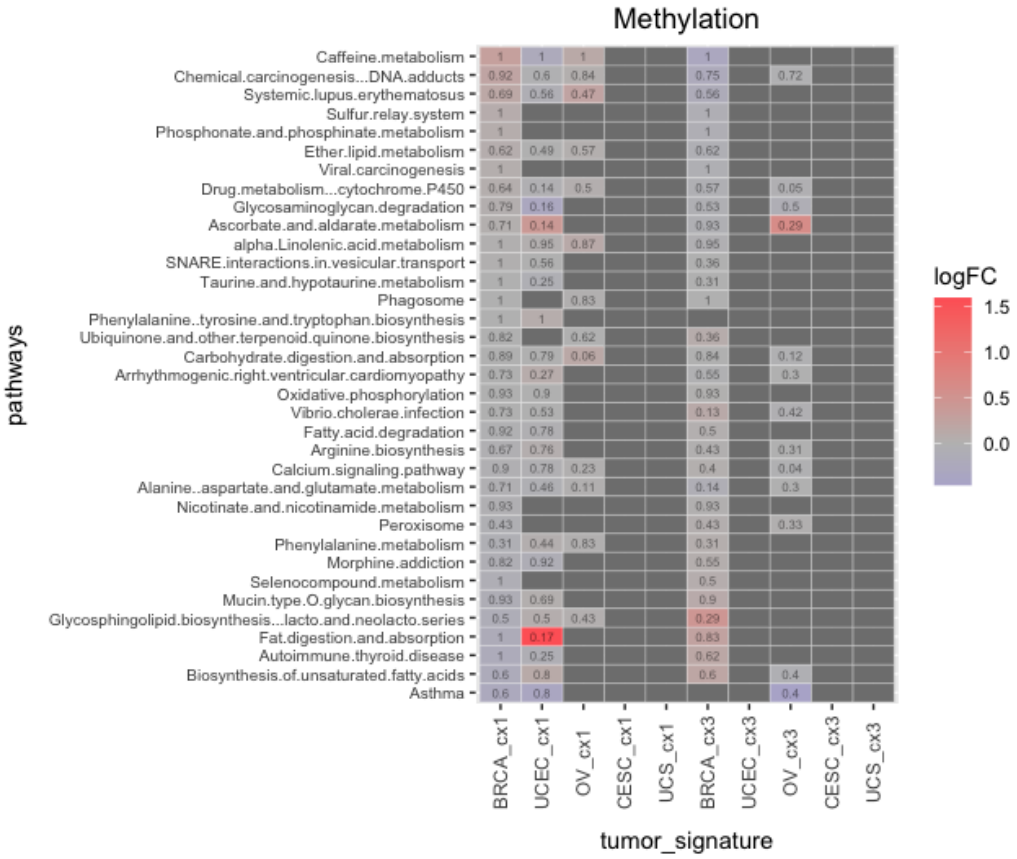


Figure 7: Heatmap showing the top 10 KEGG pathways for primary impact and absolute value of logFC detected across analyses at methylation level. The color of cells corresponds to the median logFC for the pathway, while the number inside cells is the primary impact. Grey cells correspond to pathways that were not detected as significant or having a primary impact equal to 0.

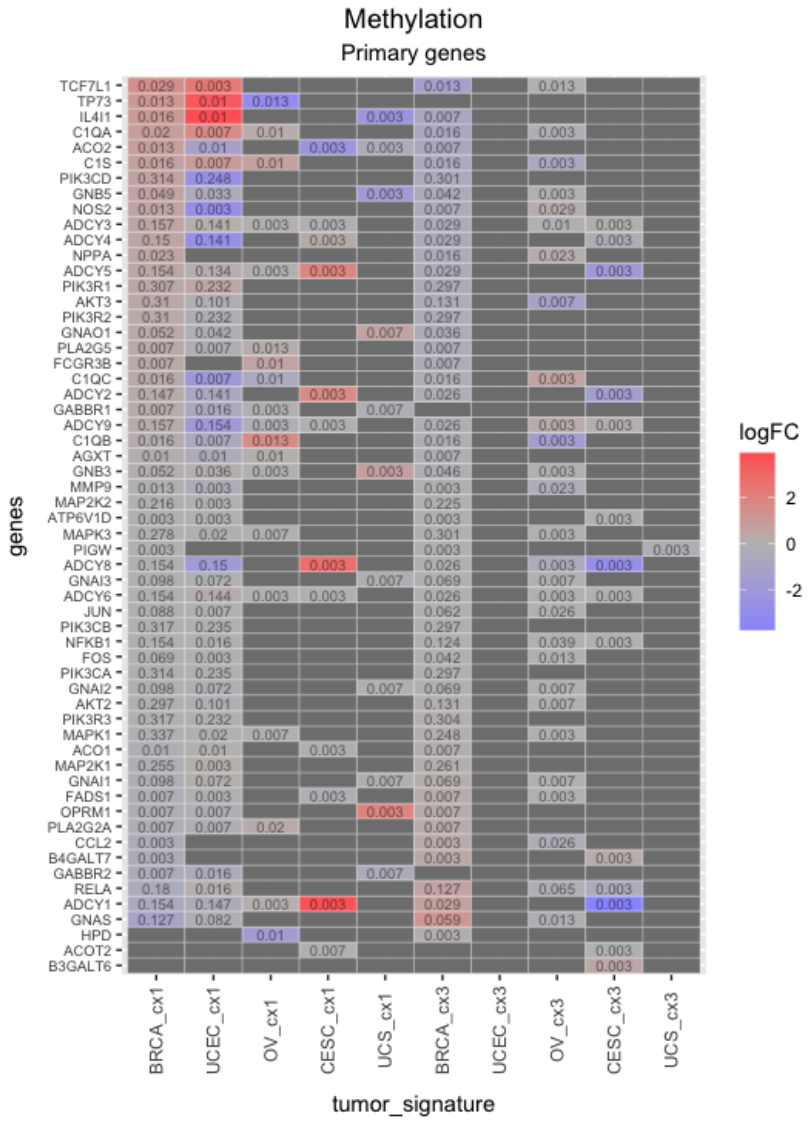


Figure 8: Heatmap showing the top 10 primary genes from KEGG pathways for relevance and logFC detected across analyses at methylation level. The color of each cell corresponds to the logFC for the gene, while the number inside cells is the relevance. Grey cells correspond to genes that were not detected as primary in any pathway (relevance = 0).

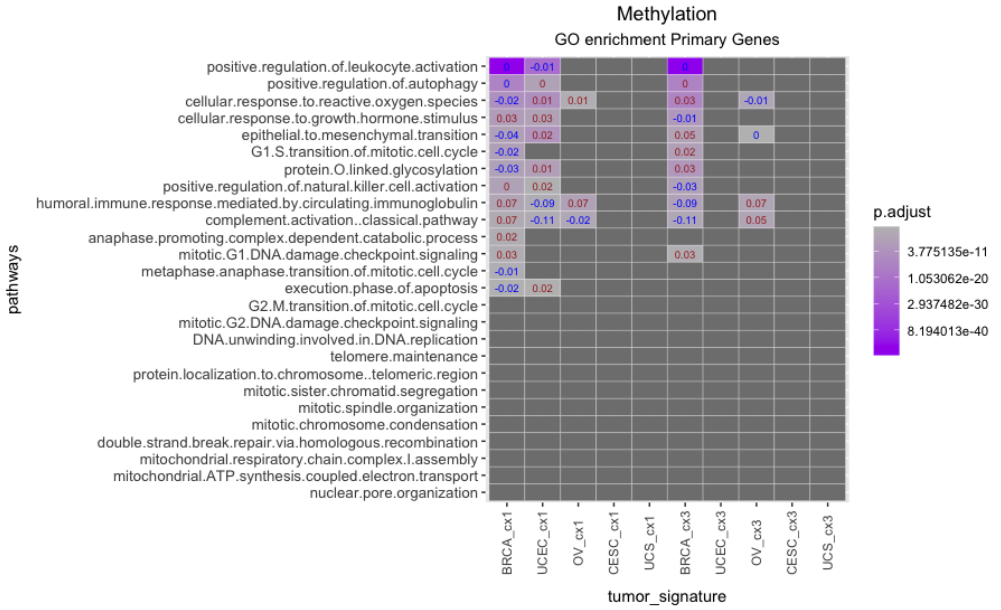


Figure 9: Heatmap showing a selection of significant GO terms enriched on primary genes detected across methylation analyses. Each cell is colored according to the value of adjusted p-value. Inside each cell the median logFC of primary genes belonging to the gene set is depicted, colored in red or blue when it is positive or negative respectively. Dark grey cells represent non-significant GO terms.

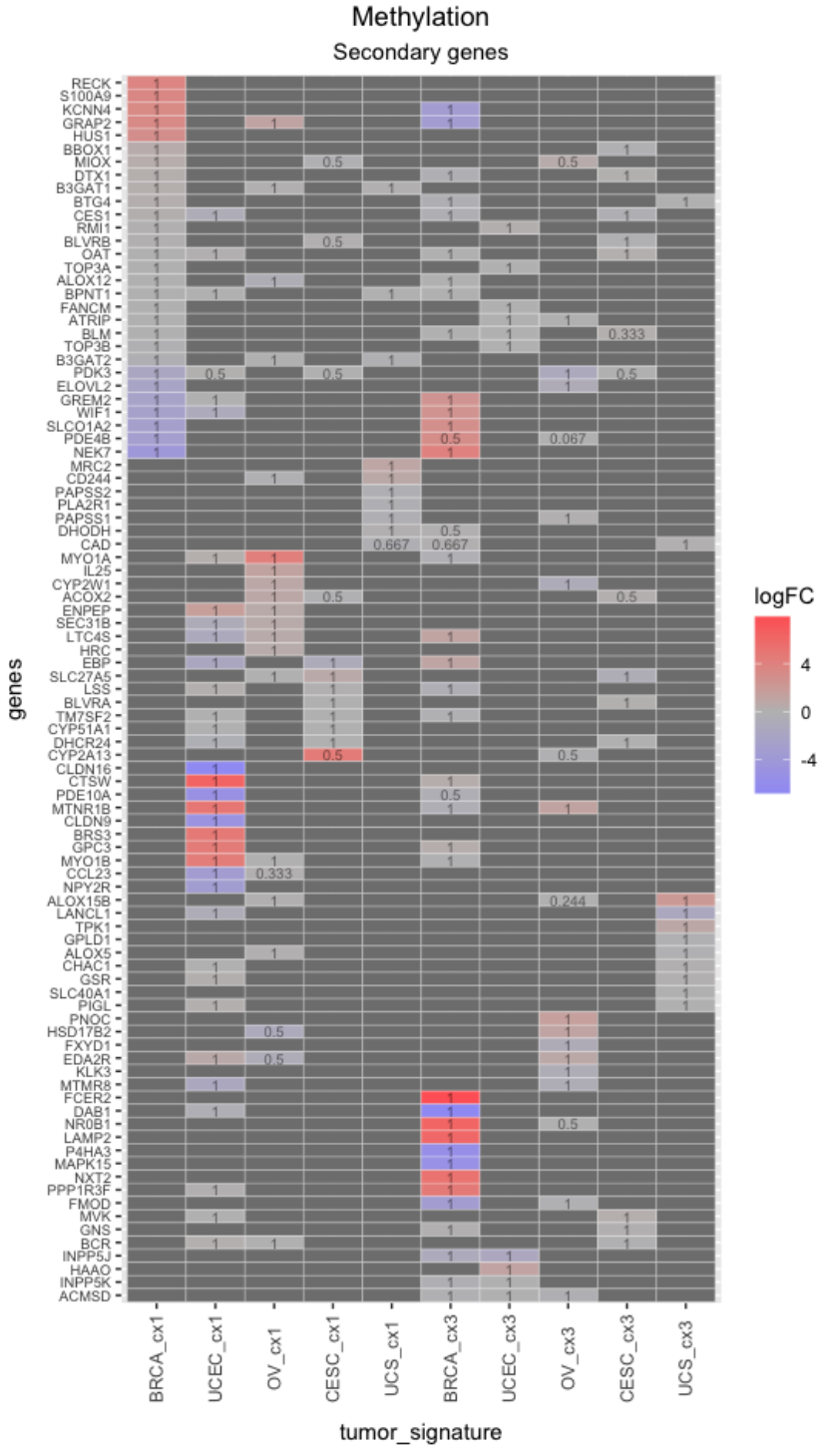


Figure 10: Heatmap showing the top 10 secondary genes for total impact and absolute value of logFC detected across methylation analyses. Cells are colored according to the logFC of the gene, inside each cell the total impact is depicted.

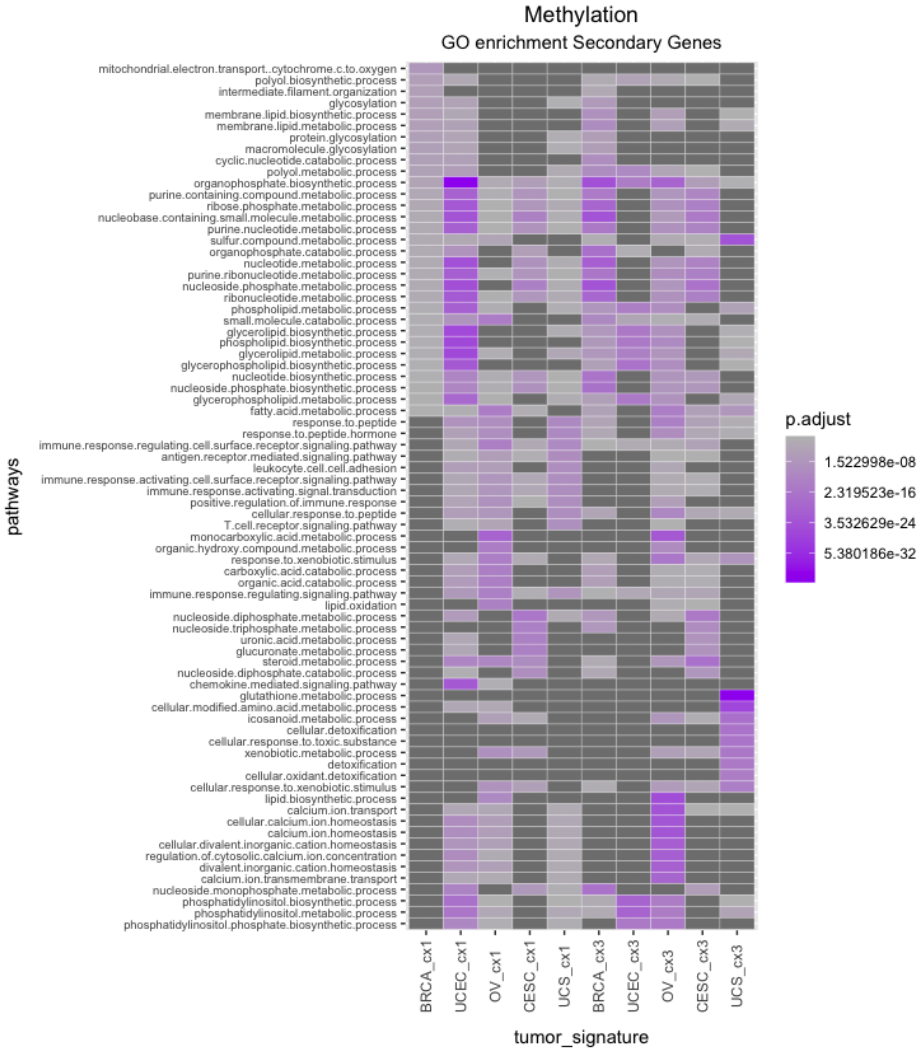
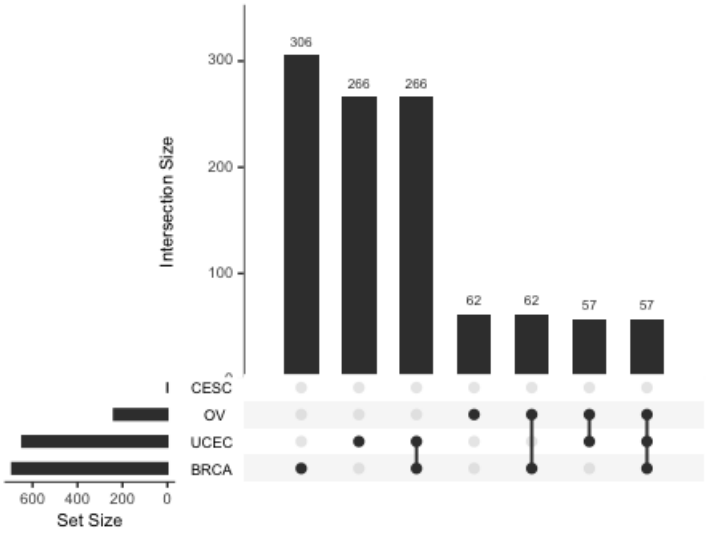
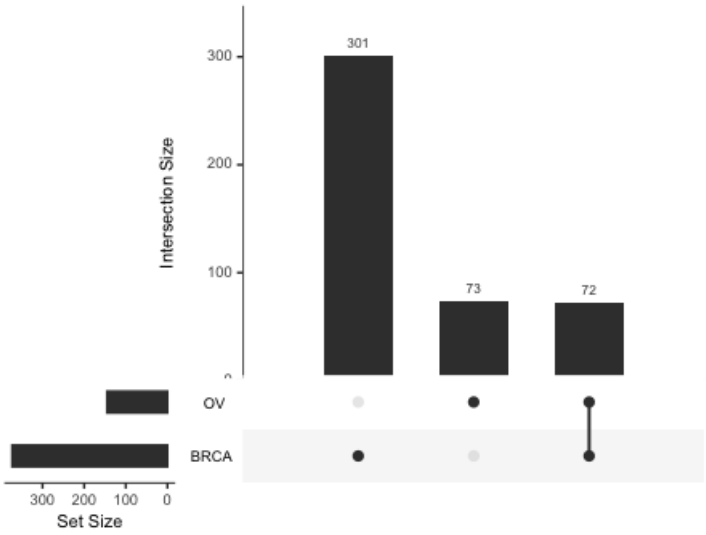


Figure 11: Heatmap showing the top 10 GO terms enriched on secondary genes. Cells are colored according to the adjusted p-value.



(a) Intersection sizes of KEGG pathways detected as significant across each combination of tumors at methylation level for analyses on CX1.



(b) Intersection sizes of KEGG pathways detected as significant across each combination of tumors at methylation level for analyses on CX3.

Figure 12

A.1.3 Expression Vs Methylation: anticorrelated genes

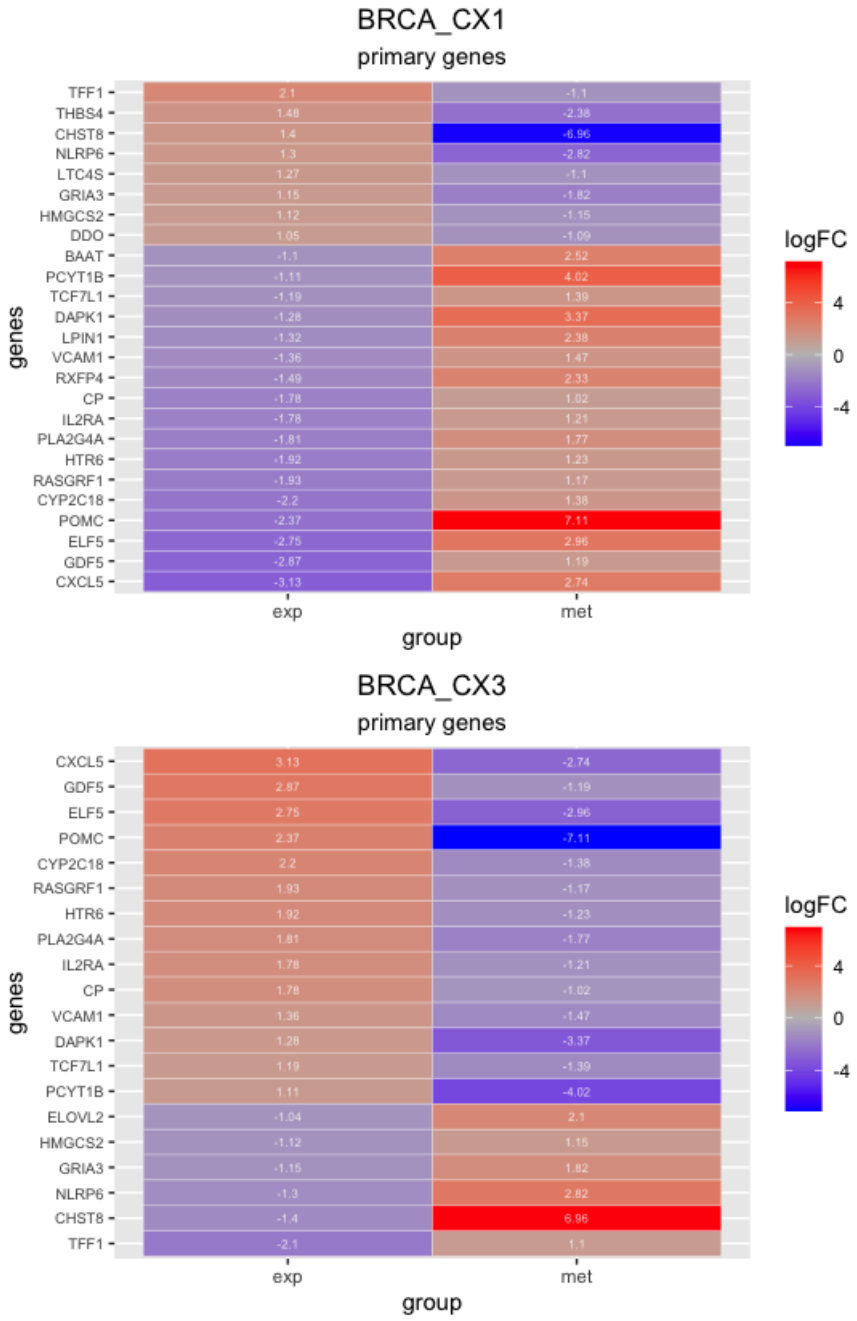


Figure 13: Heatmaps showing anticorrelated primary genes with logFC > 1 or logFC $<$ -1 for both expression and methylation, relative to BRCA analyses on CX1 and CX3. Cells are colored according to the logFC, that is also indicated inside.

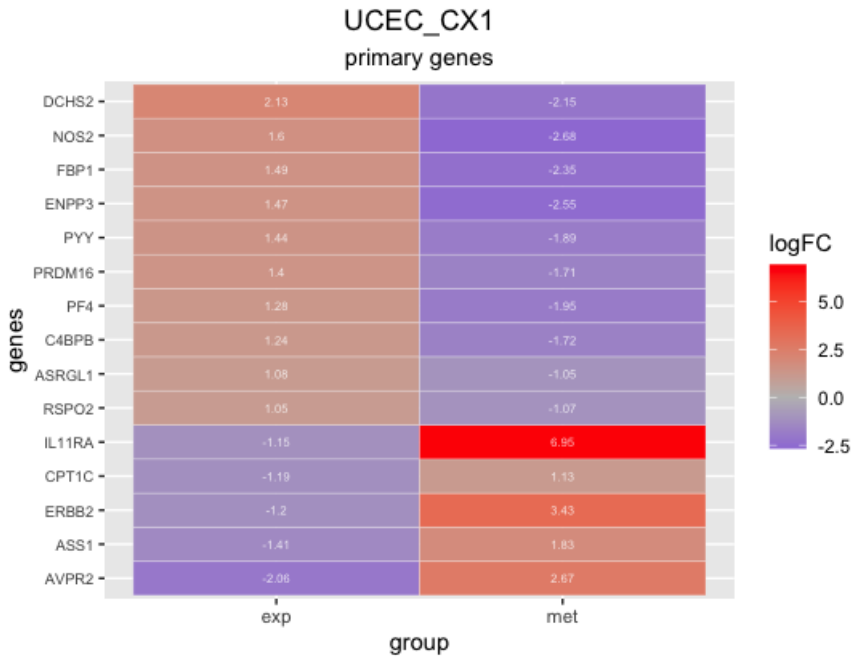


Figure 14: Heatmaps showing anticorrelated primary genes with logFC > 1 or logFC $<$ -1 for both expression and methylation, relative to UCEC analysis on CX1. Cells are colored according to the logFC, that is also indicated inside.

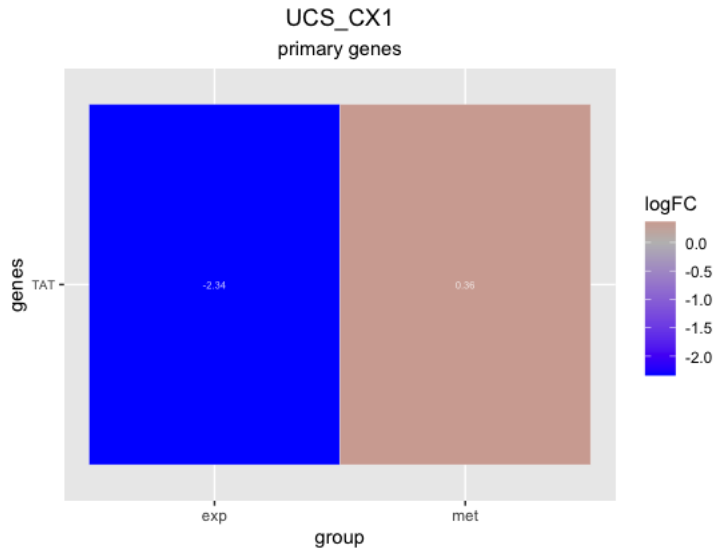


Figure 15: Heatmaps showing anticorrelated primary genes with logFC > 0.1 or logFC $<$ -0.1 for both expression and methylation, relative to UCS analysis on CX1. Cells are colored according to the logFC, that is also indicated inside.

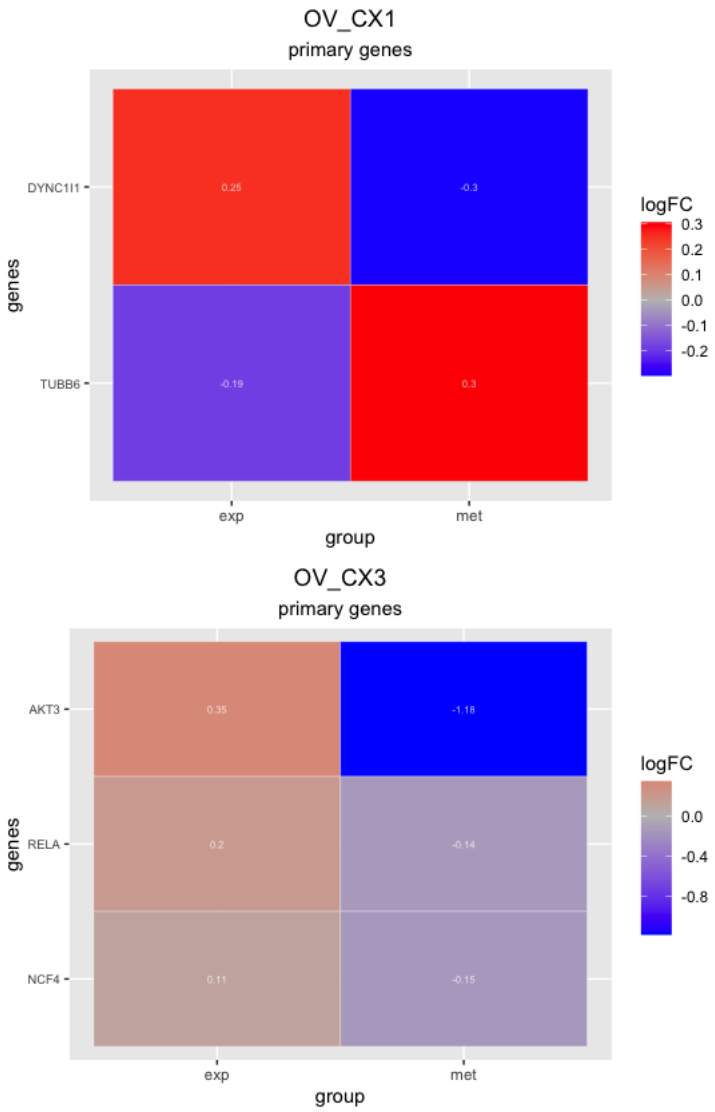


Figure 16: Heatmaps showing anticorrelated primary genes with logFC $>$ 0.1 or logFC $<$ -0.1 for both expression and methylation, relative to OV analyses on CX1 and CX3. Cells are colored according to the logFC, that is also indicated inside.

KEGG (primary genes)

		Expression	Methylation	Expression ∩ Methylation	Anticorrelated
CX1	BRCA	4311	3959	3503	1532
	UCEC	1658	1808	886	445
	OV	259	324	46	24
	CESC	67	15	0	0
	UCS	33	29	5	3
CX3	BRCA	4054	2940	2582	1143
	UCEC	60	0	0	0
	OV	521	270	38	18
	CESC	9	19	0	0
	UCS	0	1	0	0

Table 1: Number of primary genes detected in expression and methylation analyses on KEGG, the amount of genes that were found in both omics and the number of anticorrelated genes which show opposite logFC signs between expression and methylation.

A.2 Additional Reactome results

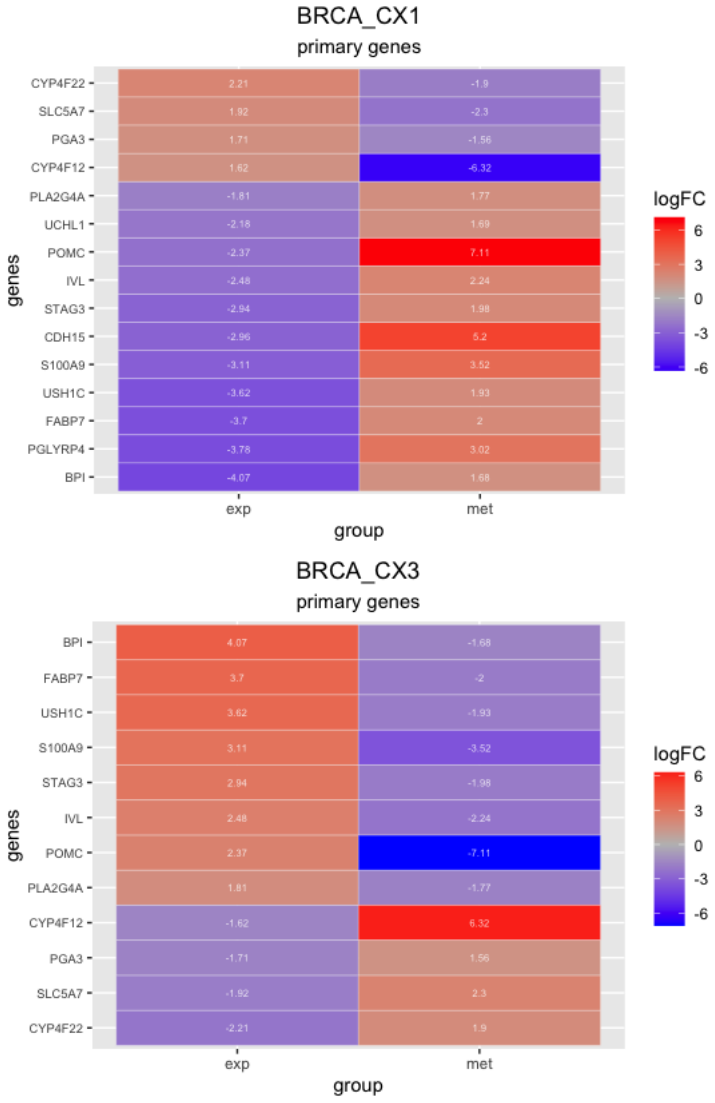


Figure 17: Heatmaps showing anticorrelated primary genes with logFC > 1.5 or logFC < -1.5 for both expression and methylation, relative to BRCA analyses on CX1 and CX3. Cells are colored according to the logFC, that is also indicated inside.

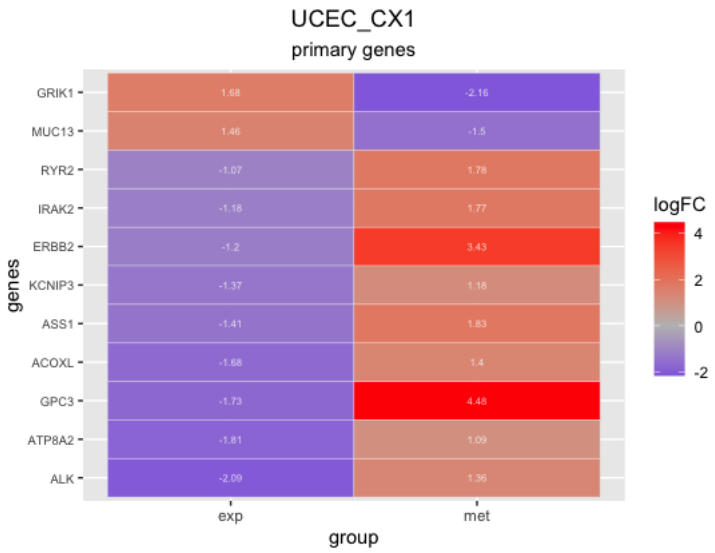


Figure 18: Heatmaps showing anticorrelated primary genes with logFC > 1 or logFC < -1 for both expression and methylation, relative to UCEC analyses on CX1. Cells are colored according to the logFC, that is also indicated inside.

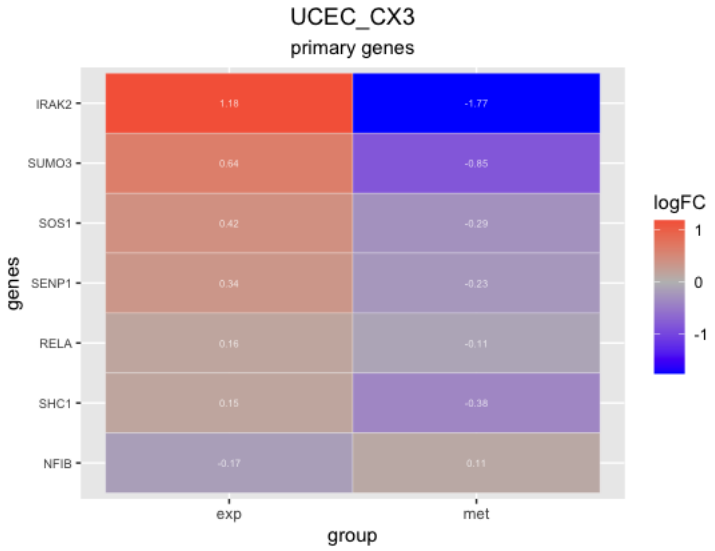


Figure 19: Heatmaps showing anticorrelated primary genes with logFC > 0.1 or logFC < -0.1 for both expression and methylation, relative to UCEC analyses on CX3. Cells are colored according to the logFC, that is also indicated inside.

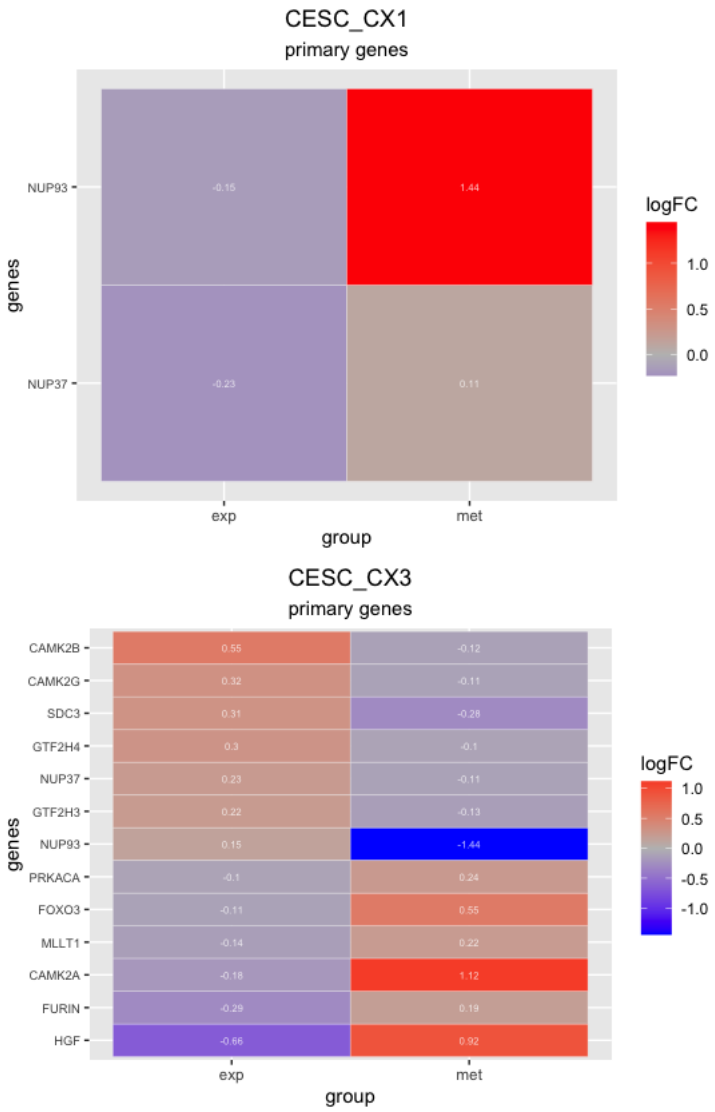


Figure 20: Heatmaps showing anticorrelated primary genes with logFC > 0.1 or logFC < -0.1 for both expression and methylation, relative to CESC analyses on CX1 and CX3. Cells are colored according to the logFC, that is also indicated inside.