CONCILIUM, Vol. 23, Nº 4, 2023
DOI: 10.53660/CLM-1015-23C35
ISSN: 0010-5236

C-Gemis: A computational tool for gene expression data analysis for gastric cancer

C-Gemis: Uma ferramenta computacional para análise de dados de expressão gênica de câncer gástrico

Received: 2023-02-10 | Accepted: 2023-03-20 | Published: 2023-03-30

Marcos Vinicius Rossetto
ORCID: https://orcid.org/0000-0002-6310-5913
University of Caxias do Sul, Brazil
E-mail: rossettomarcos@gmail.com

Paola Dutra da Rosa
ORCID: https://orcid.org/0000-0002-7357-5062
University of Caxias do Sul, Brazil
E-mail:pdrosa@ucs.br

Ivaine Thais Sauthier Sartor
ORCID: https://orcid.org/0000-0002-8775-6622
University of Caxias do Sul, Brazil
E-mail: ivaine.sauthier@gmail.com

Scheila de Avila e Silva
ORCID: https://orcid.org/0000-0002-3472-3907
University of Caxias do Sul, Brazil
sasilva6@ucs.br

ABSTRACT

Background: Computational tools dedicated to the analysis of transcripts microarray and RNA-seq can provide an instrument to search biomarkers related to diagnosis and prognosis in different neoplasia. This process is carried out by automatizing the computational process that allows the exploration, visualization, and analysis of gene expression data. **Objective:** The present paper describes a new tool named C-Gemis for gene expression data analysis. **Methods:** C-Gemis is an online and free computation tool that explores differential gene expression and survival analysis with visualization of results. **Results:** C-Gemis optimizes the search for Gastric Cancer (GC) biomarkers in available data from public databases and stands out in usability, objectivity, and easy-to-understand graphics presentation. The results are presented considering Laurén's, WHO, and TCGA molecular classification. The tool is available at the website: www.cgemis.com.br. **Conclusions:** C-Gemis provides an easy way to automate data analysis of microarray and RNA-seq. The following steps incorporate other types of cancer, reaching a high detail related to cancer classifications and subclassifications.

Keywords: Gastric cancer; TCGA; GEO; survival analysis.

RESUMO

Antecedentes: Ferramentas computacionais dedicadas à análise de transcritos microarray e RNA-seq podem fornecer um instrumento para pesquisar biomarcadores relacionados ao diagnóstico e prognóstico em diferentes neoplasias. Isso é feito pela automatização do processo computacional que permite a exploração, visualização e análise dos dados de expressão gênica. **Objetivo:** O presente artigo descreve uma nova ferramenta chamada C-Gemis para análise de dados de expressão gênica. **Métodos:** C-Gemis é uma ferramenta computacional online e gratuita que permite a exploração da expressão diferencial de genes e análise de sobrevivência e com visualização dos resultados. **Resultados:** O C-Gemis otimiza a busca de biomarcadores de Câncer Gástrico (CG) em dados disponíveis em bancos de dados públicos e se destaca pela usabilidade, objetividade e apresentação gráfica de fácil compreensão. Os resultados são apresentados considerando a classificação de Laurén, a classificação da OMS e a classificação molecular TCGA. A ferramenta está disponível no site: www.cgemis.com.br. **Conclusões:** C-Gemis fornece uma maneira fácil de automatizar a análise de dados de microarray e RNA-seq. Os próximos passos são a incorporação de outros tipos de câncer, atingindo um alto nível de detalhamento relacionado às classificações e subclassificações do câncer.

Palavras-chave: Câncer gástrico;; TCGA; GEO; análise de sobrevivência.

INTRODUCTION

The term cancer is applied to a broad and heterogeneous class of pathologies related to the accumulation of genetic alterations that trigger uncontrolled cellular proliferation. Consequently, tumor progression and its maintenance are observed. Among human neoplasia, gastric cancer (GC) is recognized as the fifth most diagnosed malignancy and the fourth leading cause of cancer-related deaths (SUNG et al., 2021).

The GC classification is based on tumor histology and has been complemented with gene expression analysis (GOLUB et al., 1999; WAN, 2015). In particular, messenger RNA (mRNA) is critical to the progression and maintenance of tumor cells. In addition, it can reflect cellular phenotypes (CRISTESCU et al., 2015). In this context, tools dedicated to analyzing transcripts from large-scale techniques, such as microarray and RNA-seq, can provide an instrument to search biomarkers related to diagnosis and prognosis in different neoplasia (BROADHEAD et al., 2010; D'ANGELO et al., 2014).

The data generated by microarray and RNA-seq techniques are available in public repositories, such as Gene Expression Omnibus (GEO) (CLOUGH E BARRETT 2016), ArrayExpress (SARKANS et al., 2021), and the consortium The Cancer Genome Atlas (TCGA). Data analysis requires knowledge about the type of data generated from molecular techniques and computational knowledge about methods and programming languages to manage the information appropriately. However, in some cases, researchers need to gain molecular biology or computer science expertise to deal with data exploration and analysis. Therefore, searching for biomarkers in GC requires a computational tool dedicated to automating the computational process that allows the exploration, visualization, and analysis of gene expression data.

The definition of biomarker proposed by the World Health Organization (WHO) is "any substance, structure or process that can be measured in the body or its products and that can influence or predict the incidence of outcomes or diseases." The first step in identifying biomarkers is to understand the pathophysiology of the disease and its associated factors. Therefore, the potential biomarker must be correlated with clinical variables, e.g., histological classification and tumor extension.

There are available tools for gene expression analysis, such as GEO2R (BARRETT et al., 2013), GDAC Firehose, and cBioPortal (CERAMI et al., 2012; GAO et al., 2013). GEO2R performs comparative analyses from data available in the GEO database to identify differentially expressed genes of two or more groups of samples. In its execution, GEO2R allows the visualization of gene expression values distribution per sample. In addition, it offers some correction options to control the occurrence of false positives for multiple comparisons, such as the Benjamini-Hochberg test. Finally, at the end of the execution, GEO2R provides a graphical presentation of the gene expression profile, and the R-script used to execute the entire process. Nonetheless, its execution is laborious and timely expansive since the location of the code of the target gene in the index file is not standard across all different platforms that perform the microarray technique.

The GDAC Firehose is a tool developed by Broad Institute that systematically analyzes TCGA data. The tool stores about 40 terabytes of data collected from the TCGA database, updated twice a month. The platform contains a series of standardized pipelines for genomic and transcriptomic analysis. The GDAC provides standardized and updated results at different levels, such as revised data, analysis of results, and user-friendly reports for researchers. The computing environment is available either online or locally installed. However, sometimes the online availability of the tool is off in its use. Another limitation is the choice of genes since it is impossible to analyze more than one gene at a time (YANG et al., 2015).

The cBioPortal CERAMI et al. 2012; GAO et al. 2013) is a web portal that visualizes multidimensional data from genomic analysis. It contains data from CCLE (Cancer Cell Line Encyclopedia) and TCGA (WU et al., 2019). On the cBioPortal website, the researchers can explore molecular change patterns by comparing multiple studies or summarizing the relevant variations in an individual tumor sample. Another feature of the tool is the exploration of metabolic pathways, survival analysis, and the possibility of data download.

(YANG et al. 2015). Recently, the cBioPortal website became an open-source software, which means it is possible to improve the source code (WU et al., 2019). Nonetheless, the portal does not perform a detailed histotype analysis, such as the cancer subclassification.

In this way, in order to contribute to gene expression analysis, starting from data acquisition and processing until results visualization, the present work describes a new tool named C-Gemis. Currently, the first version of C-Gemis is targeted only for gastric cancer, but soon, the tool will be expanded to other types of cancer. The C-Gemis data source is TCGA and GEO, and this tool aims to contribute to the process of biomarkers search, including the gene expression data analysis in a tumor.

METHODS

C-Gemis is a free computation tool dedicated to the exploration of differential gene expression and survival analysis and with visualization of results. The architecture of C-Gemis is divided into two parts, the graphical user interface and the data processing layer. The user interface was developed using Angular 12. The data processing layer was developed in the R language, and it is responsible for the acquisition, preprocessing, normalization, and data analysis using specialized function packages, according to (Table 1).

Table 1 – R function packages

Libraries	**Functions**
agricolae (MENDIBURU 2014)	To perform the statistical difference between groups
survival (HAIBE-KAINS et al., 2008; SINNOTT e CAI 2016); survcomp (SCHRÖDER et al., 2011)	To perform survival analysis
GEOquery (DAVIS 2007)	To download GEO datasets
TCGAbiolinks (COLAPRICO et al., 2016); oligo (CARVALHO E IRIZARRY 2010); affy (GAUTIER et al., 2004)	To preprocess and normalize the microarray and RNA-seq data
hgu133plus2.db (CARLSON 2017); pd.huex.1.0.st.v2 (BENILTON 2017).	To map the microarray probes

RESULTS

C- Gemis is a web user interface compatible with major browsers and operating systems, including Safari, Internet Explorer, Firefox, and Chrome, using Linux, Windows, or Apple iOS. The tool can be accessed at the following URL

To access C-Gemis, the user does not need to perform any registration. The tool interface is organized at the following levels: (i) the home page that presents an overview of the tool and its objectives; (ii) an analysis of the data obtained from the TCGA database; (iii) a help page that shows the operation and some frequently asked questions; (i) about us page, which contains publications and the research group that developed C-Gemis.

To perform differential gene expression analysis, the user must choose between TCGA or GEO databases since each database requires specific packages, as there are differences in data format and study design, according to (Figure 1) and (Figure 2).

Figure 1. Workflow of GEO analysis.

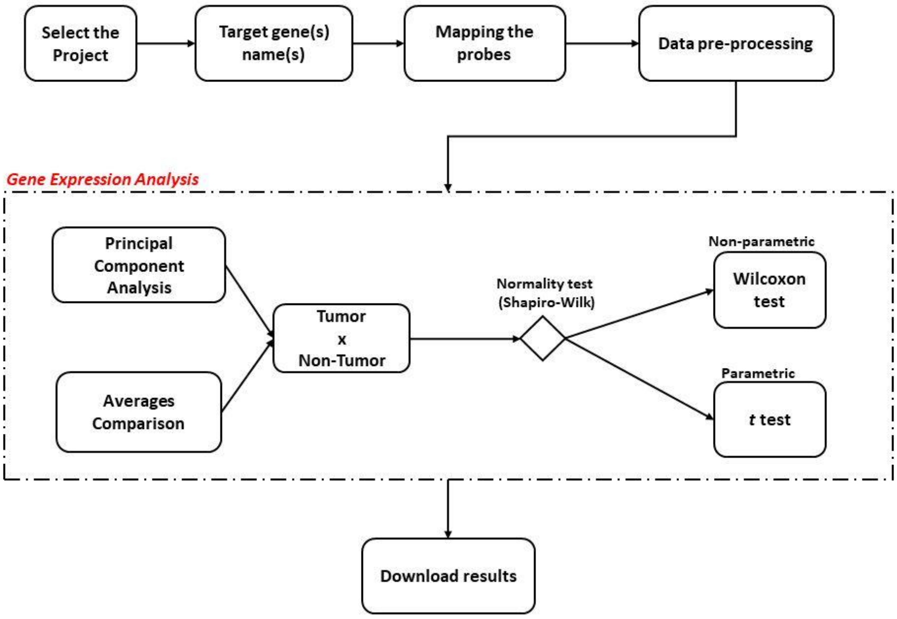
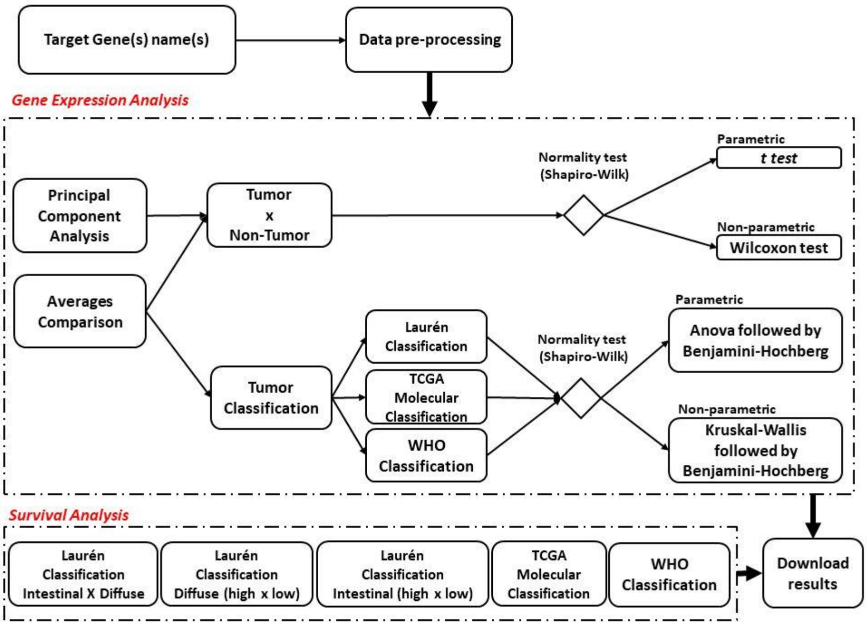


Figure 2. Workflow of TCGA analysis.


Target Gene(s) name(s)
Data pre-processing
Gene Expression Analysis
Principal Component Analysis
Averages Comparison
Tumor x Non-Tumor
Normality test (Shapiro-Wilk)
Parametric
t test
Non-parametric
Wilcoxon test
Tumor Classification
Laurén Classification
TCGA Molecular Classification
WHO Classification
Normality test (Shapiro-Wilk)
Parametric
Anova followed by Benjamini-Hochberg
Non-parametric
Kruskal-Wallis followed by Benjamini-Hochberg
Survival Analysis
Laurén Classification Intestinal X Diffuse
Laurén Classification Diffuse (high x low)
Laurén Classification Intestinal (high x low)
TCGA Molecular Classification
WHO Classification
Download results

Independently of the database source, the user must enter the name of the gene(s) of interest in the format mapped in HUGO Genes and runs the analysis. When all scripts are executed, the user interface will show all results as image files that can be downloaded. If the data comes from the GEO repository, the preprocessing and normalization are made automatically by C-Gemis. Afterward, statistical analysis and mapping of the probes are performed according to the microarray platform only for the genes inserted by the user. On the other hand, if the data comes from the TCGA, it is possible to present some additional results, such as survival information and the comparison of gene expression according to the different subclassifications of gastric cancer.

DISCUSSION

Principal Results

The development of C-Gemis aims to optimize the search for GC biomarkers in available data from public databases. C-Gemis stands out in usability, objectivity, and easy-to-understand graphics presentation. The histological classifications in C-Gemis are Laurén's classification (LAUREN 1965), WHO's, and TCGA molecular classification. C-Gemis also brings the survival analysis, classified according to Laurén, WHO, and TCGA. In addition, the tool provides survival analysis comparing the median expression levels of the gene of interest according to Lauren's classification.

Limitations

Currently, the first version of C-Gemis is targeted only for gastric cancer, but shortly, the tool will be expanded to other types of cancer.

Comparison with Prior Work

Compared to the current tools available (GDCA Firehouse, GEO2R, cBioPortal), C-Gemis presents the following attributes that can be highlighted: further details on the classifications of the gastric cancer subtypes and the possibility of exploration of prognostic biomarkers using the gene expression.

The GDAC Firehouse tool analyzes the gene of interest among all tumor types available and presents a general graphical visualization of the expression profile. However, GDAC Firehouse does not analyze in detail the tumor subtypes. Instead, the tool links molecular changes with clinical characteristics. However, it summarizes the results with a significant association, which searches for a gene of interest more costly. In addition, GDAC Firehouse compares gene expression with other studies. Although the completeness of cBioPortal, it does not present both gene expression and survival analysis based on the classifications of cancer subtypes. Therefore, C-Gemis explores the full potential of the collected data.

C-Gemis executes the appropriate statistical test, whereas GEO2R requires the researcher to have the analytical knowledge to conduct gene expression comparisons since statistical analysis is required to compare the groups. C-Gemis also facilitate gene selection since it uses the HUGO gene annotation, whereas GEO2R involves the selection of the probe related to the gene of interest. This information is obtained exclusively by searching the probe code in a separate file downloaded within the gene expression data. However, this information is not easy to find since there are several probes for the same gene, and, in some cases, there is no indication of the best probe that should be used to represent a gene. Considering this issue, C-Gemis automatically performs the mapping of the probes related to the gene of interest, and the best probe is informed. Otherwise, when no best probe exists, C-Gemis performs an average of all related probes.

CONCLUSIONS

C-Gemis provides an easy way to automate data analysis of microarray and RNA- seq. Therefore, the investigation and the search for diagnostic and prognostic biomarkers in gastric cancer gain another instrument. The following steps incorporate other types of cancer, reaching a high detail related to cancer classifications and subclassifications.

ACKNOWLEDGEMENTS

We are grateful for the support received from Universidade de Caxias do Sul (UCS) and the Coordernação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) by the Ph.D. scholarship.

Conflicts of Interest

None declared.

Abbreviations

GC: Gastric Cancer

mRNA: Messenger Ribonucleic Acid

GEO: Gene Expression Omnibus

TCGA: The Cancer Genome Atlas

WHO: World Health Organization

CCLE: Cancer Cell Line Encyclopedia

REFERÊNCIAS

BARRETT, T. et al. **NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids** Res. 2013 Jan;41(Database issue):D991-5. doi: 10.1093/nar/gks1193. Epub 2012 Nov 27. PMID: 23193258; PMCID: PMC3531084.

BENILTON, C. pd.mogene.2.0.st [Internet]. **Bioconductor**; 2017 [cited 2022 Jan 24]. Available from: https://bioconductor.org/packages/pd.mogene.2.0.st.

BROAD INSTITUTE TCGA GENOME DATA ANALYSIS CENTER. **Analysis Overview for Stomach Adenocarcinoma (Primary solid tumor cohort) -** Jan 15 2014 [Internet]. Broad Institute of MIT and Harvard; 2014. Available from: http://gdac.broadinstitute.org/runs/analyses__2014_01_15/reports/cancer/STAD-TP/index.html [acessed 2022-01-24]

BROADHEAD, ML.; CLARK, JC.; DASS, CR.; CHOONG, PF. **Microarray: an instrument for cancer surgeons of the future?** ANZ J Surg. 2010 Jul-Aug;80(7-8):531-6. doi: 10.1111/j.1445-2197.2010.05379.x. PMID: 20795968.

CANCER GENOME ATLAS RESEARCH NETWORK. **Comprehensive molecular characterization of gastric adenocarcinoma**. Nature. 2014 Sep 11;513(7517):202-9. doi: 10.1038/nature13480. Epub 2014 Jul 23. PMID: 25079317; PMCID: PMC4170219.

CARLSON, M. hgu133plus2.db [Internet]. **Bioconductor**; 2017 [cited 2022 Jan 24]. Available from: https://bioconductor.org/packages/hgu133plus2.db.

CARVALHO, BS. Irizarry RA. **A framework for oligonucleotide microarray preprocessing. Bioinformatics.** 2010 Oct 1;26(19):2363-7. doi: 10.1093/bioinformatics/btq431. Epub 2010 Aug 5. PMID: 20688976; PMCID: PMC2944196.

CERAMI, E. et al. **The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data**. Cancer Discov. 2012 May;2(5):401-4. doi: 10.1158/2159-8290.CD-12-0095. Erratum in: Cancer Discov. 2012 Oct;2(10):960. PMID: 22588877; PMCID: PMC3956037.

CGEMIS. Available from: www.cgemis.com.br

CLOUGH, E.; BARRETT, T. **The Gene Expression Omnibus Database**. Methods Mol Biol. 2016;1418:93-110. doi: 10.1007/978-1-4939-3578-9_5. PMID: 27008011; PMCID: PMC4944384.

COLAPRICO, A. et al. **TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data.** Nucleic Acids Res. 2016 May 5;44(8):e71. doi: 10.1093/nar/gkv1507. Epub 2015 Dec 23. PMID: 26704973; PMCID: PMC4856967.

CRISTESCU, R. et al. **Molecular analysis of gastric cancer identifies subtypes associated with distinct clinical outcomes.** Nat Med. 2015 May;21(5):449-56. doi: 10.1038/nm.3850. Epub 2015 Apr 20. PMID: 25894828.

D'ANGELO, G.; DI RIENZO, T.; OJETTI, V. **Microarray analysis in gastric cancer: a review**. World J Gastroenterol. 2014 Sep 14;20(34):11972-6. doi: 10.3748/wjg.v20.i34.11972. PMID: 25232233; PMCID: PMC4161784.

DAVIS S.; MELTZER, PS. **GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor.** Bioinformatics. 2007 Jul 15;23(14):1846-7. doi: 10.1093/bioinformatics/btm254. Epub 2007 May 12. PMID: 17496320.

DE MENDIBURU, F. (2014). **Agricolae: statistical procedures for agricultural research.** R package version, 1(1).

GAO, J. et al. **Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal.** Sci Signal. 2013 Apr 2;6(269):pl1. doi: 10.1126/scisignal.2004088. PMID: 23550210; PMCID: PMC4160307.

GAUTIER L.; COPE, L.; BOLSTAD, BM.; IRIZARRY, RA. af**fy--analysis of Affymetrix GeneChip data at the probe level.** Bioinformatics. 2004 Feb 12;20(3):307-15. doi: 10.1093/bioinformatics/btg405. PMID: 14960456.

GOLUB, T. et al. **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.** Science. 1999 Oct 15;286(5439):531-7. doi: 10.1126/science.286.5439.531. PMID: 10521349.

HAIBE-KAINS, B. et al . **A comparative study of survival models for breast cancer prognostication based on microarray data: does a single gene beat them all?** Bioinformatics. 2008 Oct 1;24(19):2200-8. doi: 10.1093/bioinformatics/btn374. Epub 2008 Jul 17. PMID: 18635567; PMCID: PMC2553442.

LAUREN, P. The two histological main types of gastric carcinoma: diffuse and so-called intestinal-type carcinoma. An attempt at a histo-clinical classification. Acta Pathol Microbiol Scand. 1965;64:31-49. doi: 10.1111/apm.1965.64.1.31. PMID: 14320675.

NAGTEGAAL, ID. et al. **WHO Classification of Tumours Editorial Board. The 2019 WHO classification of tumours of the digestive system.** Histopathology. 2020 Jan;76(2):182-188. doi: 10.1111/his.13975. Epub 2019 Nov 13. PMID: 31433515; PMCID: PMC7003895.

SARKANS, U. et al. **From ArrayExpress to BioStudies.** Nucleic Acids Res. 2021 Jan 8;49(D1):D1502-D1506. doi: 10.1093/nar/gkaa1062. PMID: 33211879; PMCID: PMC7778911.

SCHRÖDER, MS. et al. **Survcomp: an R/Bioconductor package for performance assessment and comparison of survival models.** Bioinformatics. 2011 Nov 15;27(22):3206-8. doi: 10.1093/bioinformatics/btr511. Epub 2011 Sep 7. PMID: 21903630; PMCID: PMC3208391.

SINNOTT, JA.; CAI, T. **Inference for survival prediction under the regularized Cox model. Biostatistics.** 2016 Oct;17(4):692-707. doi: 10.1093/biostatistics/kxw016. Epub 2016 Apr 22. PMID: 27107008; PMCID: PMC5031946.

SUNG, H. et al. **Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries.** CA Cancer J Clin. 2021 May;71(3):209-249. doi: 10.3322/caac.21660. PMID: 33538338

TCGA. About TCGA. URL www.cancergenome.nih.gov/abouttcga [acessed 2022-01-24]

WAN, C.; LI, J. **Synthesis of well-dispersed magnetic CoFe2O4 nanoparticles in cellulose aerogels via a facile oxidative co-precipitation method.** Carbohydr Polym. 2015 Dec 10;134:144-50. doi: 10.1016/j.carbpol.2015.07.083. Epub 2015 Aug 5. PMID: 26428110.

WORLD HEALTH ORGANIZATION (WHO), International Programme on Chemical Safety. Biomarkers in Risk Assessment: Validity and Validation, 2001.

WU, P. et al. **Integration and Analysis of CPTAC Proteomics Data in the Context of Cancer Genomics in the cBioPortal.** Mol Cell Proteomics. 2019 Sep;18(9):1893-1898. doi: 10.1074/mcp.TIR119.001673. Epub 2019 Jul 15. PMID: 31308250; PMCID: PMC6731080.

YANG, Y. et al. **Databases and web tools for cancer genomics study. Genomics Proteomics Bioinformatics.** 2015 Feb;13(1):46-50. doi: 10.1016/j.gpb.2015.01.005. Epub 2015 Feb 21. Erratum in: Genomics Proteomics Bioinformatics. 2015 Jun;13(3):202-203. PMID: 25707591; PMCID: PMC4411507.