Activity-regulated cytoskeleton-associated protein (ARC) and Cancer mutation

Germán Cano Amaro

ARC protein plays a major role on the brain, development and functioning, thus, its importance is notable.The goal of this study was to determine possible relations between ARC protein and cancer. Taking into consideration the protein structure and the cancer distribution over the body tissue the study was done by comparing different databases as the Catalogue Of Somatic Mutation In Cancer, the National Institute of Cancer or Universal Protein bank. It was found that lung, skin and colon cancer along with others, could be related with mutations on mainly the C-lobe of the protein.

INTRODUCTION

Synaptic plasticity is fundamental on the interaction between neurons and how they control any aspect of the brain since it determine emotional responses, cognitive flexibility and memory formation. As a mRNA binding protein with important role on the neurons nucleus, ARC protein has been recently proposed as a key regulator of long-term synaptic plasticity and depression [1] ,hence, it plays a major role on our brain development and life.

Although there are still open questions related to how ARC protein works or interact with other proteins, it can be already noticed how important ARC is. It is then of high interest for different scientific fields such as medicine, biology or psychology to understand and define its main characteristics and properties as well as its possible relations, direct or indirect, with diseases. This study was focused on the relation between ARC protein and Cancer since other studies already show a possible relation between this protein and different types of cancer [2], [3], [4] . However, this does not mean that it does not have implications in other diseases (mostly related with the brain) as it is for example Alzheimer, schizophrenia or anxiety disorder [5],[6] ,[7] .

The sequence of ARC protein has been well defined as well as it structure which is mainly composed by a C-terminal (278-360) which enclose a C-lobe (278-364), a N-terminal (46-154) and a N-lobe (211-277). Since it is a protein that take part on neuron interaction it main tissue distribution is on the brain. Mutations of ARC protein have been studied and identified along the entire sequence , still, no direct association with any disease has been observed. All we could do was infer indirect consequences from this mutations.

Even though no variation was directly associated with any disease it is of interest to know which of thus variations on the sequence could lead potentially to cancer and how it could be related to the structure in which the mutation happen as well as the type of cancer.

The approach to answer this question was to identify the variations on the sequence and the structure of the protein in which they occurs as well as if there is any registered relation between the mutation and any type of cancer.

METHODS

Four different databases were used to obtain the data for this analysis, the main ones were the National Institute of Cancer (NIH) [8] and the Catalogue Of Somatic Mutation In Cancer (COSMIC) database [9] used to extract information about the possible mutations on the ARC protein sequence, as well as the cancer each mutation was related to. The first one was composed by 12981 samples from different studies with 117 mutations observed along all the studies while the former by 40519 samples with 345 mutations. In addition, the SWISS-MODEL repository [10] was used to determine the different structures of the protein as well as their position on the protein sequence. All the data gathered from these databases was supported and compared with the Uniprot (Universal Protein) database [11]. For a deeper analysis on the protein structure WebLogo was used for generating the sequence logo of each protein structure [12] and the order of each structure was calculated by Predictions of Intrinsically Unstructured Proteins (IUPRED) [13] .

Both COSMIC and NIH databases offer a distribution of the tissues where cancer associated mutations were found. At this point a distribution was made by simply counting the frequency with which each tissues appeared on the databases. In the case of the COSMIC dataset they already provide the percentage of mutations per tissue taking into account the size of the sample they used to study each tissue individually. On the other hand NIH database do not provide such percentage since they use data from different studies. The distribution for NIH dataset was made then taking into account the total number of mutated samples found along all the studies. Whereas a list of identified mutations appear in the COSMIC and NIH databases, the former classifies each of them using the ensemble variant effect predictor (VEP) [14] . Considering the mutations marked with a high VEP impact a comparison in terms of sequence position and type of mutation were made between databases as well as an analysis of the tissue distribution for those mutations. Finally a general study over their distribution along the different structures of the protein was performed regarding also the properties of each structure.

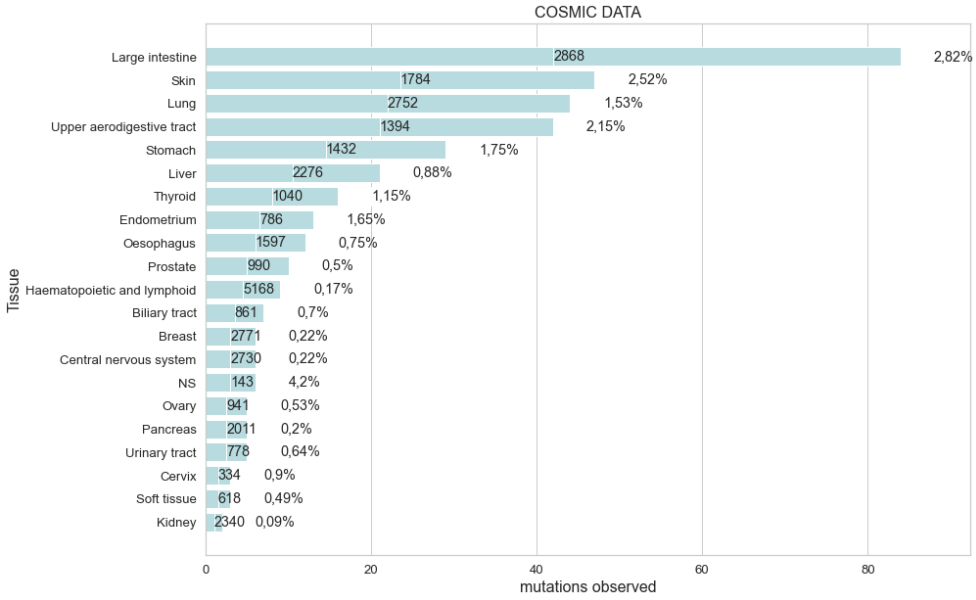

FIG. 1. Tissue distribution of cancer related with ARC protein within COSMIC database. The numbers on each bar represent the number of samples studied for each tissue and the percentage of mutations found with respect the total samples of each tissue.

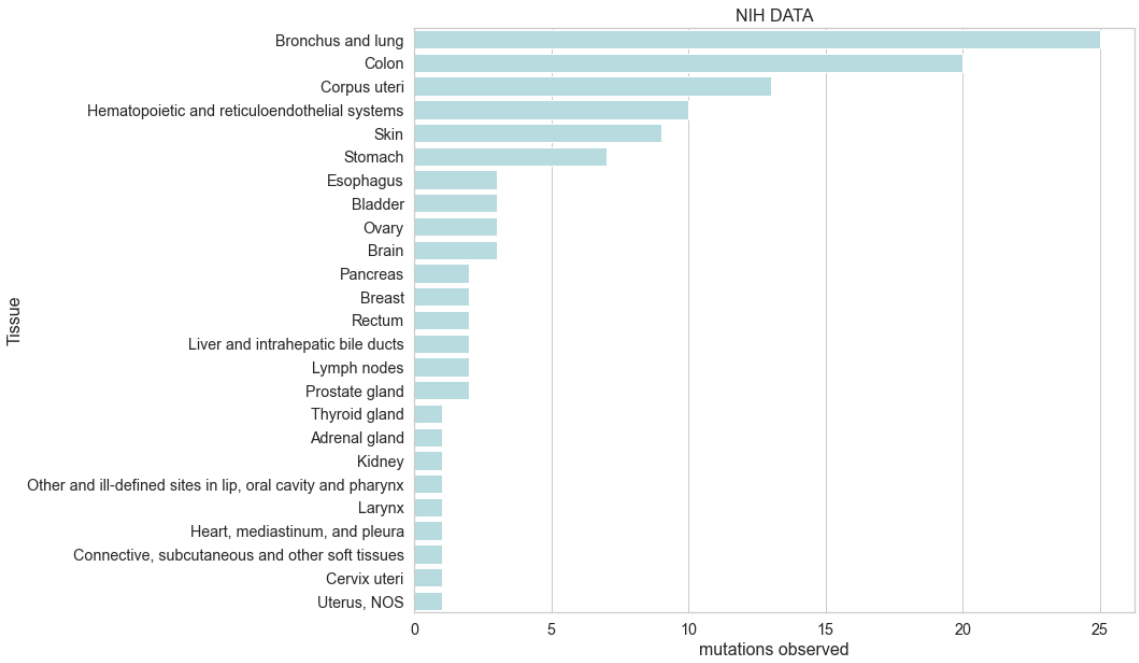

FIG. 2. Tissue distribution of cancer related with ARC protein within NIH database.

RESULTS

ARC protein and tissue distribution of cancer.

The first study was conducted on how different types of cancer could be related with mutations of ARC protein independently on the mutation position, this led to establish a distribution of those cancers over different tissues. Filtering NIH and COSMIC databases allowed to count the number of times that each tissue shows a cancer related with ARC within their samples. Furthermore on COSMIC database the percentage of mutated samples on each tissue was represented.

On COSMIC database, as shown in 1 it could be observed how large intestine and skin followed by the upper

digestive tract have the highest percentage of cancer related with ARC even if they are not the tissues with more samples studied. On the other hand kidney and lymphoid tissue have the lowest percentage with the same or more samples studied than the previous tissues. NS tissue has the highest percentage but since it represent non specific tissue it was not considered for the analysis. This contrast with NIH database, where the highest number of mutations were observed on bronchus, lung, colon and corpus uteri tissue 2 .

Taking into consideration the differences between the two databases, as the number of total samples or the tissues studied, it could be observed how mutations on ARC protein could apparently be related with cancer on lung, skin and large intestine (colon). However, without the number of samples studied on each tissue for NIH database it was not possible to make a real comparison between the percentage that each tissue represent. This is the reason why tissues, as hematopoietic or uterus, show high differences on each database. In addition, considering the COSMIC database, which has the higher number of samples, and given the percentage of mutations related with cancer on each tissue, which are all under 3% . No direct association between cancer and mutation could be established for any specific tissue.

High VEP impact on ARC protein mutations.

Once determined the relation between cancer on different tissues and mutations of ARC protein a deeper study was performed over all the mutations. This way, with support of the previous study, it was possible to determine if those mutations with higher VEP impact parameter were related to the cancers most frequently associated with ARC. VEP impact parameter was mainly shown on NIH database, thus the position of each mutation in the sequence was extracted from there, however, since UNIPROT database also offers information on this parameter it was used as comparison with NIH data. For more precision, the appearance of each mutation was compared over NIH, COSMIC and UNIPROT in addition to the types of mutations over the whole sequence.

There are several types of mutations along the ARC protein sequence as can be observed in I. Missense substitutions, which have a probability of leading to a cancer, are predominant followed by synonymous substitutions , which are harmless. From all of them, ten mutations with high VEP impact parameter were found along the ARC protein sequence. Associated with different cancers on different tissues their position on the sequence has been shown on II. When comparing with the other databases, differences in both the value of the parameter and the mutation itself were appreciated. The first remarkable difference was that some mutations and their position differs from each database, there was no mutation found in positions 165 and 303 in the COSMIC database while 313 did not appear on UNIPROT. In addition the VEP impact parameter was not specified in UNIPROT for positions 165, 303 and 318. Last difference found was that the mutation on 318 was marked as $G > A$ on COSMIC. UNIPROT shows the two possibilities on this position and another mutation $G > T$ marked as moderated impact on 310.

Type of mutation	Number of samples	% from total
Missense subtitutions	208	60.29
Nonsense subtitutions	17	4.93
Synonymous subtitutions	89	25.8
Frameshift insertion	5	1.45
Frameshift deletion	3	0.87
Other (not specified)	26	7.54

TABLE I. Total number of mutations and their structure distribution.

Position	Tissue
165	Brain, Bronchius, lung, kidney, pancreas, uterus
238	Uterus
303	Breast, colon, ovary, peritoneum
310	Bronchus and lung
313	Ovary
318	Hematopoietic and reticuloendothelial systems
323	Skin
330	Colon, rectosigmoid junction
338	Uterus
341	Colon, rectosigmoid junction
354	Bronchus and lung

TABLE II. High VEP impact mutations.

Among the tissues that shows cancer derived from high VEP impact parameter mutations, the ones with higher percentage of cancer related with ARC, as lung, colon or skin were found. It was also shown how in tissues with low percentage of relation such as kidney and pancreas, high impact mutations were observed. Although, the tissue with more high VEP impact related mutations was the feminine reproductive system which, as seen before, is one of the tissues where more mutation were observed in NIH dataset and it has one of the highest percentage on COSMIC. However the differences between databases and the size of study sample for each suggest that this results could not be definitive.

Cancer distribution over ARC protein structures.

Further analysis was conducted on how the mutations were distributed over the different structures of ARC protein as well as if there were more high impact mutations in one structures than others. By determining the exact distribution of the structures on the sequence with the SWISS-MODEL database the number of mutations corresponding to each structure were counted in both COSMIC and NIH.

Structure	C-terminal(C-lobe)	N-lobe	N-terminal	Rest
NIH	30	15	16	17
COSMIC	77	45	60	70

TABLE III. Total number of mutations and their structure distribution.

As can be observed on III both datasets show a similar distribution. It has to be mentioned that taking into consideration the previous study on high impact mutations, the C-terminal presents seven high impact mutations while the N-lobe one and the N-terminal does not presents any. Considering the total number of mutated samples of each database a comparison between the mutations observed on each structure by each dataset was done 3 . It was observed how COSMIC and NIH show similar tendencies. However while on NIH the C-lobe shows a significant difference on the number of mutation with respect the other structures on the COSMIC database all the mutations were more equally distributed.

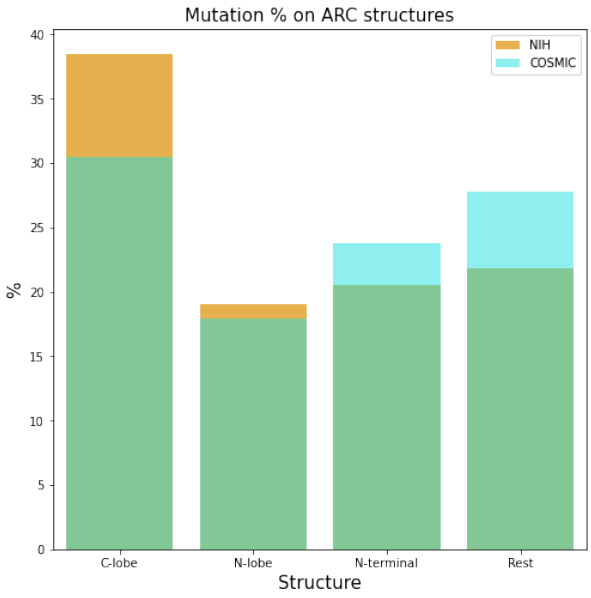

FIG. 3. Distribution of the mutations over the different structures of ARC protein based on the percentage they represent on NIH and COSMIC databases.

Results suggest that mutations on C-lobe have a stronger relation with cancer than mutations on the other structures which is supported by the fact that most of the high impact mutations occurs there. The difference that both databases shows could be explained by the number of samples that each database took for the study. Since COSMIC used more samples, the statistics of its data should be more accurate, hence even if mutations on C-lobe seems to be more related with cancer, mutations on N-terminal shows also relation with non high impact cancers.

ARC structure order and cancer.

The last analysis consisted on a deeper study over the protein structure order and how it could be correlated with the distribution of cancer along the structures. Comparing the sequence logo of each structure with their order led to the establishment of a relation between how ordered was each structure and the distribution of cancer on them. Firstly for the N-terminal by simply looking at

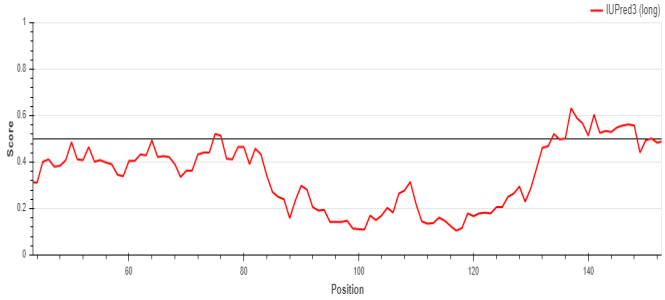

FIG. 4. Order score along ARC N-terminal sequence calculated by IUPRED.

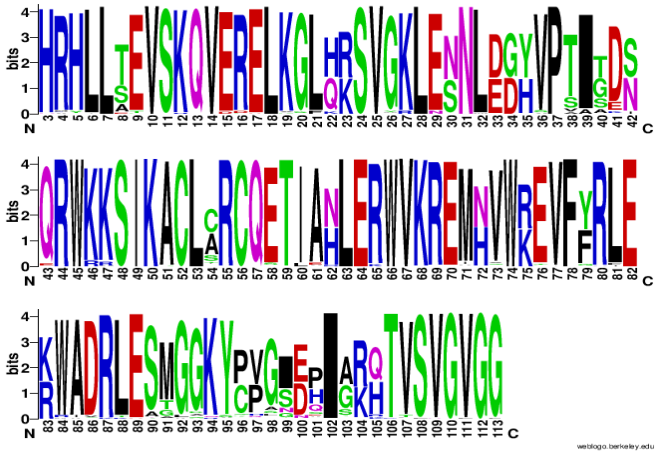

FIG. 5. Sequence logo of N-terminal from ARC protein

its sequence logo 5 we observe how overall all the letters has the same height and few of them are stacked which

mean that in general the sequence remains stable with little changes. This fact is supported by the order analysis of the structure observed in 4 where positions above the marked score are classified as disordered, and the ones under the marked score as ordered. It can be noticed how nearly the end of the sequence the score is established as disordered which coincide with the last part of the sequence logo where more stacked letters with less height could be observed. On the other hand the C-lobe showed some differences with respect the N-terminal. Whereas its sequence logo 7 looks similar to the one of N-terminal it could be appreciated how there are more stacked letters with less height mainly at the beginning and the end of the sequence. Again, the order analysis of this structure 6 revealed how at the beginning and at the end of the sequence the structure is classified mostly as disordered.

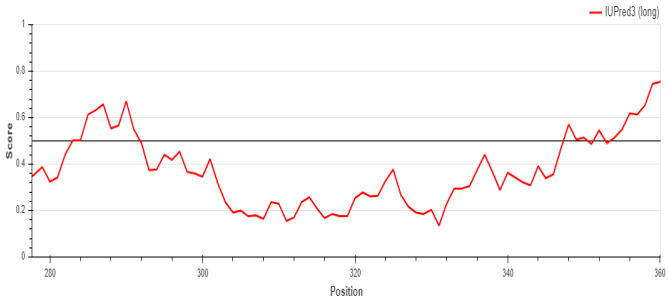

FIG. 6. Order score along ARC C-lobe sequence calculated by IUPRED.

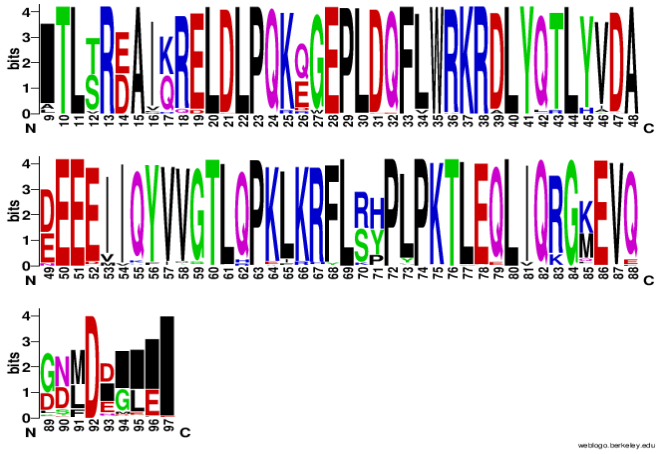

FIG. 7. Sequence logo of C-lobe from ARC protein.

This observations present the C-lobe as a more disordered structure than the N-terminal, which would be considered ordered, within the ARC protein. Ordered structures are known to be functionally more important for the protein so they are more conserved during evolution, disordered ones, instead, are functionally less important and hence more likely to suffer mutations. Since C-lobe is a more disordered structure it presents the higher number of mutations whereas the N-terminal as a more ordered structure present less which was already observed in 3. This, could also explain why the high impact mutations appear mostly on C-lobe, if more mutations occurs there the probability of one of them being of high impact is also higher. Thus, it could be said that N-terminal is functionally a more important structure for the protein than C-lobe.

CONCLUSIONS

In this study, the aim was to establish a relationship between sequence mutations of ARC protein and cancer, including the tissue distribution of the cancer. the structure distribution of the mutations and it relation with the order of each structure . In that line it was found that cancers on lung, skin and large intestine have, a priori, higher relation with ARC protein mutations as suggested on [2]. On the other hand, not significant relation could be established with brain cancer as [4] proposed. From all the mutations, ten were classified with high VEP impact parameter. The majority of them were related with cancer on the feminine reproductive system, however they were also identified mostly on lung and large intestine. The distribution of the mutation over the protein structure showed how mutations mostly occurs on the C-lobe, this was supported by the sequence logo and the order analysis of each structure which showed the C-lobe to be a disordered structure and the N-terminal ordered as proposed on [15].

Summarizing, mutations on ARC protein mainly show up on its disordered structures as C-lobe. This could potentially lead to different types of cancer between which; lung, rectum, skin and uterus seem to be more probable. Even though different databases were used, the poor size on the sample of study of each, made stating any definitive conclusion or correlation between ARC mutations and cancer not rigorous.

REFERENCES

[1] O. Nikolaienko, S. Patil, M. S. Eriksen, and C. R. Bramham, "Arc protein: a flexible hub for synaptic plasticity and cognition", in Seminars in cell & developmental biology, Vol. 77 (Elsevier, 2018), pp. 33–42.

[2] U. G. Bhat and A. L. Gartel, "Differential sensitivity of human colon cancer cell lines to the nucleoside analogs arc and drb", International journal of cancer **122**, 1426–1429 (2008).

[3] G. C. Gobe, K. L. Ng, D. M. Small, D. A. Vesey, D. W. Johnson, H. Samaratunga, K. Oliver, S. Wood, J. L. Barclay, R. Rajandram, et al., "Decreased apoptosis repressor with caspase recruitment domain confers resistance to sunitinib in renal cell carcinoma through alternate angiogenesis

pathways", Biochemical and biophysical research communications **473**, 47–53 (2016).

[4]Q. Wang, A. Li, H. Wang, and J. Wang, "Knockdown of apoptosis repressor with caspase recruitment domain (arc) increases the sensitivity of human glioma cell line u251mg to vm-26", International Journal of Clinical and Experimental Pathology **5**, 555 (2012).

[5]S. Landgren, M. von Otter, M. S. Palmér, C. Zetterström, S. Nilsson, I. Skoog, D. R. Gustafson, L. Minthon, A. Wallin, N. Andreasen, et al., "A novel arc gene polymorphism is associated with reduced risk of alzheimer's disease", Journal of neural transmission **119**, 833–842 (2012).

[6]Y. Chen, S. Bang, M. F. McMullen, H. Kazi, K. Talbot, M.-X. Ho, G. Carlson, S. E. Arnold, W.-Y. Ong, and S. F. Kim, "Neuronal activity-induced sterol regulatory element binding protein-1 (srebp1) is disrupted in dysbindin-null mice—potential link to cognitive impairment in schizophrenia", Molecular neurobiology **54**, 1699–1709 (2017).

[7]R. D. Penrod, J. Kumar, L. N. Smith, D. McCalley, T. B. Nentwig, B. W. Hughes, G. M. Barry, K. Glover, M. Taniguchi, and C. W. Cowan, "Activity-regulated cytoskeleton-associated protein (arc/arg3. 1) regulates anxiety-and novelty-related behaviors", Genes, Brain and Behavior **18**, e12561 (2019).

[8]N. I. of Mental Health, *ARC*, (2022) https://portal.gdc.cancer.gov/genes/ENSG00000198576 (visited on 05/18/2022).

[9]COSMIC, *ARC GENE*, (2022) https://cancer.sanger.ac.uk/cosmic/gene/analysis?ln=ARC#tissue (visited on 05/18/2022).

[10]S. Model, *ARC Structure*, (2022) https://swissmodel.expasy.org/repository/uniprot/Q7LC44?range=206-364&template=6gse.1.A (visited on 05/18/2022).

[11]Uniprot, *Q7LC44 · ARC_HUMAN*, (2022) https://beta.uniprot.org/uniprotkb/Q7LC44/entry (visited on 05/18/2022).

[12]G. E. Crooks, G. Hon, J.-M. Chandonia, and S. E. Brenner, "Weblogo: a sequence logo generator", Genome research **14**, 1188–1190 (2004).

[13]G. Erdős, M. Pajkos, and Z. Dosztányi, "Iupred3: prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of evolutionary conservation", Nucleic acids research **49**, W297–W303 (2021).

[14]D. Martin, M. C. Abba, A. A. Molinolo, L. Vitale-Cross, Z. Wang, M. Zaida, N. C. Delic, Y. Samuels, J. G. Lyons, and J. S. Gutkind, "The head and neck cancer cell oncogenome: a platform for the development of precision molecular therapies", Oncotarget **5**, 8906 (2014).

[15]M. Boldridge, J. Shimabukuro, K. Nakamatsu, C. Won, C. Jansen, H. Turner, and L. Wang, "Characterization of the c-terminal tail of the arc protein", Plos one **15**, e0239870 (2020).

[16]S. Heikaus, T. Kempf, C. Mahotka, H. E. Gabbert, and U. Ramp, "Caspase-8 and its inhibitors in rccs in vivo: the prominent role of arc", Apoptosis **13**, 938–949 (2008).

[17]W. McLaren, L. Gil, S. E. Hunt, H. S. Riat, G. R. Ritchie, A. Thormann, P. Flicek, and F. Cunningham, "The ensembl variant effect predictor", Genome biology **17**, 1–14 (2016).

[18]E. I. Hallin, C. R. Bramham, and P. Kursula, "Structural properties and peptide ligand binding of the capsid homology domains of human arc", Biochemistry and Biophysics Reports **26**, 100975 (2021).

[19]J. G. Tate, S. Bamford, H. C. Jubb, Z. Sondka, D. M. Beare, N. Bindal, H. Boutselakis, C. G. Cole, C. Creatore, E. Dawson, et al., "Cosmic: the catalogue of somatic mutations in cancer", Nucleic acids research **47**, D941–D947 (2019).

[20]"Uniprot: the universal protein knowledgebase in 2021", Nucleic acids research **49**, D480–D489 (2021).