**META-ANALYSIS AND STRUCTURE PREDICTION OF DYSREGULATED LNCRNAs IN SELECT CANCERS FOLLOWED BY SUB-CELLULAR LOCALIZATION STUDIES**

***Submitted by:***

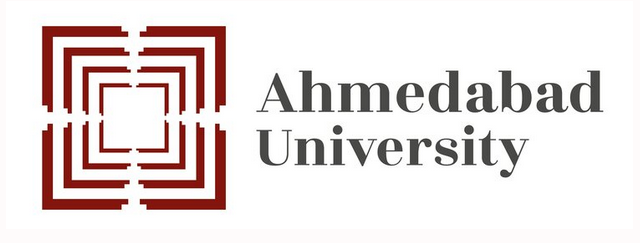
**DRASHTI RIKEN SHAH**

**AU1621043**

**Session: 2016-2021**

**A Dissertation submitted to the Faculty of Ahmedabad University in Partial Fulfilment of the requirement for the Degree of**

**Integrated Masters of Life Sciences**

****

**Biological and Life Sciences**

**School of Arts and Sciences**

**Ahmedabad University**

**Ahmedabad-380009**

**Gujarat**

**June 2021**

**CERTIFICATE**

This is to certify that the dissertation entitled **“**META-ANALYSIS AND STRUCTURE PREDICTION OF DYSREGULATED LNCRNAs IN SELECT CANCERS FOLLOWED BY SUB-CELLULAR LOCALIZATION STUDIES**”** submitted by **Ms. Drashti Riken Shah, 1621043** , to the Biological and Life Sciences, School of Arts and Sciences, Ahmedabad University, Ahmedabad – 380009, for the degree of **Integrated Master of Science in Life Sciences** is his/her original work, based on the results of the experiments and investigations carried out independently by him/her during the period **January – 2021** to **May- 2021** of study under the supervision of **Dr. Bhaskar Datta** at **Indian Institute of Technology, Gandhinagar.**

This is also to certify that the above said work has not been previously submitted for the award of any degree, diploma or fellowship in any Indian or Foreign University.

**Programme Chair Dean of the school of Arts and Sciences**

**CERTIFICATE OF ORIGINALITY AND CONTRIBUTION**

I hereby declare that the project titled “META-ANALYSIS AND STRUCTURE PREDICTION OF DYSREGULATED LNCRNAs IN SELECT CANCERS FOLLOWED BY SUB-CELLULAR LOCALIZATION STUDIES” is my own original work and that, to the best of my knowledge and belief, it contains no material previously completed or written by another person. This work has not been submitted as a project report at this or any other institution before, except where due acknowledgment is made in the report. Any contribution made to the report by colleagues, with whom I have worked at SAS or elsewhere, during the project, is fully acknowledged.

I also declare that the intellectual content of this report is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.

Signature:

**TABLE OF CONTENTS**

|  |  |
| --- | --- |
| ACKNOWLEDGEMENTS | 5 |
| ABSTRACT | 6 |
| LIST OF FIGURES | 7 |
| LIST OF TABLES | 8 |
| ABBREVIATIONS | 9 |
| AIM | 10 |
| INTRODUCTION | 11 |
| REVIEW OF LITERATURE | 13 |
| MATERIALS AND METHOD | 25 |
| RESULT AND DISCUSSION | 34 |
| CONCLUSION | 40 |
| FUTURE PROSPECTS | 41 |
| REFERENCES | 43 |

**ACKNOWLEDGEMENTS**

I am very thankful to Indian Institute of Technology – Gandhinagar for giving me the opportunity to broaden my horizons, grow my knowledge and help me develop new skils. I extend my gratitude towards Dr. Bhaskar Datta for his valuable guidance and supervision which helped me throughout this project. His constant encouragement and support helped me develop knowledge and insight in this new field.

I am extremely grateful towards Biological and Life Sciences, School of Arts and Sciences, Ahmedabad University for their trust in their students and allowing students to venture and explore on their own into the exciting fields of Biology.

I would extend my sincere thanks to Chinmayee Shukla and Shubham Sharma for their help in the laboratory.

Lastly, I would like to thank my friends and my family for their constant encouragement and moral support.

**ABSTRACT**

In recent years, G-quadruplexes have been implicated in a variety of studies related to certain cancers. These non-canonical structures are scattered throughout our genome and are formed in G-rich regions of DNA and RNA. They are thought to play a major role in progression and metastasis in cancers and have gained a lot of traction due to their significance in being possible therapeutic agents in these cancers. Within these, RNA G-quadruplexes occur widely in the mRNA and non-coding regions which can have a possible impact on translational regulation and chromosomal stability and integrity. Along with this, there are a number of long non-coding RNAs (non-coding RNAs with more than 200 nucleotides) which are thought to play a role in cancers. Bioinformatics tools such as QGRS Mapper and G4 hunter use their own specific algorithms to predict putative G-quadruplex forming sequences. The lncRNAs can be fed into these web-based servers and the G4 forming sequences with their G-scores listed. The aim of my study is to put these two subjects together – RNA G-quadruplexes and lncRNAs and to figure out how the information obtained can be further applied.

Key words: G-quadruplexes, long non-coding RNAs, QGRS Mapper, G4 Hunter.

**LIST OF FIGURES**

|  |  |
| --- | --- |
| FIGURE 1 | Structure of DNA and RNA |
| FIGURE 2 | Biogenesis of long non-coding RNAs |
| FIGURE 3 | Increase in the research conducted on ncRNAs |
| FIGURE 4 | Identification and characterization of lncRNAs with roles in cancer |
| FIGURE 5 | Interactions in the G4 quartet |
| FIGURE 6 | Difference between DNA and RNA G-quadruplexes |
| FIGURE 7 | Schematic of the formation of G-quadruplex DNA during normal cellular events |
| FIGURE 8 | lnc2cancer 3.0 database user interface |
| FIGURE 9 | List of lncRNAs associated with cervical cancer obtained from lnc2cancer 3.0 |
| FIGURE 10 | FASTA format of AFAP-AS1 |
| FIGURE 11 | QGRS Mapper User Interface |
| FIGURE 12 | Analysis view in QGRS Mapper |
| FIGURE 13 | List of PQS shown in QGRS Mapper |
| FIGURE 14 | Excel datasheet created for information collected from QGRS Mapper |
| FIGURE 15 | List of PQS for lncRNA TOB1-AS1 |
| FIGURE 16 | Importing sequences from NCBI Nucleotide to G4 Hunter |
| FIGURE 17 | Analysis interface of G4 Hunter |
| FIGURE 18 | G4 Hunter Analysis of lncRNA AFAP1-AS1 |
| FIGURE 19 | List of G-quadruplexes predicted in G4 Hunter |
| FIGURE 20 | Bar graph depicting the PQS distribution for cervical cancer lncRNAs |
| FIGURE 21 | Bar graph depicting the PQS distribution for lung cancer lncRNAs |
| FIGURE 22 | Plot 1 for subcellular localization in CEBPA-AS1 |
| FIGURE 23 | Plot 2 for subcellular localization in CEBPA-AS1 |

**LIST OF TABLES**

|  |  |
| --- | --- |
| TABLE 1 | Examples of lncRNAs with tumour-suppressor or oncogenic functions |
| TABLE 2 | Information on the types of PQS in cervical cancer lncRNAs and their percentage occurrence |
| TABLE 3 | Information on the G-scores of lncRNAs having 4G PQS |
| TABLE 4 | Information on the types of PQS in lung cancer lncRNAs and their percentage occurrence**.** |
| TABLE 5 | Information on the G-scores of lncRNAs having 4G PQS |
| TABLE 6 | G-scores for five lncRNAs in QGRS Mapper and G4 Hunter |

**ABBREVIATIONS**

|  |  |
| --- | --- |
| lncRNA | Long non-coding ribonucleic acid |
| G4 | G-quadruplex |
| mRNA | Messenger RNA |
| ncRNA | Non-coding ribonucleic acid |
| DNA | Deoxyribonucleic Acid |
| RNA | Ribonucleic Acid |
| QGRS | Quadruplex forming G-rich Sequences |
| NCBI | National Center for Biotechnology Information |
| PQS | Putative Quadruplex-forming Sequences |
| TCF3 | Transcription Factor 3 |
| ceRNA | Competing Endogenous RNA |
| miRNA | microRNA |
| RCI | Relative Concentration Index |
| FPKM | Fragments per kilo base per million mapped |

**AIM**

The aim of this study is to create a database for cervical and lung cancer correlating information obtained regarding lncRNAs and their putative G-quadruplex forming sequences using various bioinformatics tools.

**INTRODUCTION**

After the discovery of the double-stranded helix of DNA, it was pretty clear that this uncomplicated structure can actually adopt a multitude of configurations based on the interacting proteins around it. One of these secondary structures is a G-quadruplex. These G-rich sequences form stacks of G-tetrads, bolstered by Hoogsteen bonds and have many conformations within themselves. Shorter intervening loops and longer G-tracts confer stability to the structure. Monovalent cations present in the cavities of these quadruplexes also help in stabilizing the structure [1]. Recent studies have shown that these structures have quite a few physiological roles to play in in-vitro functions.

In a variety of organisms, G4 motifs are readily found in telomeric regions due to their high GC content and single stranded nature of the telomere overhang. We know that telomerase activity is often dysregulated in most cancers and the presence of G4 motifs in this region can be targeted for cancer therapy with the help of small molecule ligands [2]. Various studies now have also started pointing towards the importance of RNA G-quadruplexes which are mostly found in mRNA, non-coding RNAs and telomeric ends. Visual localisation using a G-quadruplex specific antibody has shown prevalence of RNA G-quadruplexes in the human cell cytoplasm [3]. The inherent structure of RNA with its 2’-hydroxyl group of the ribose sugar leads to more stability of RNA G-quadruplex than its DNA counterpart.

In the many sequences where RNA G4 is found, long non-coding RNAs is a major candidate. These are a group of non-coding RNAs with more than 200 nucleotides. In the past few years, studies have shown that these lncRNAs play a pivotal role in the progression and development of cancer. The lncRNA expression is dysregulated in cancers both transcriptionally and post-transcriptionally [4]. Some prevalent lncRNAs like HOTAIR, FENDRR, IRAIN and MALAT-1 are repeatedly mentioned in cancer research.

Due to the increasing significance of lncRNAs, a number of databases have emerged recently to put in one place the information about each lncRNAs and their expression patterns in different cancers. These databases, coupled with bioinformatics tools to predict G4- forming sequences can prove to be extremely useful.

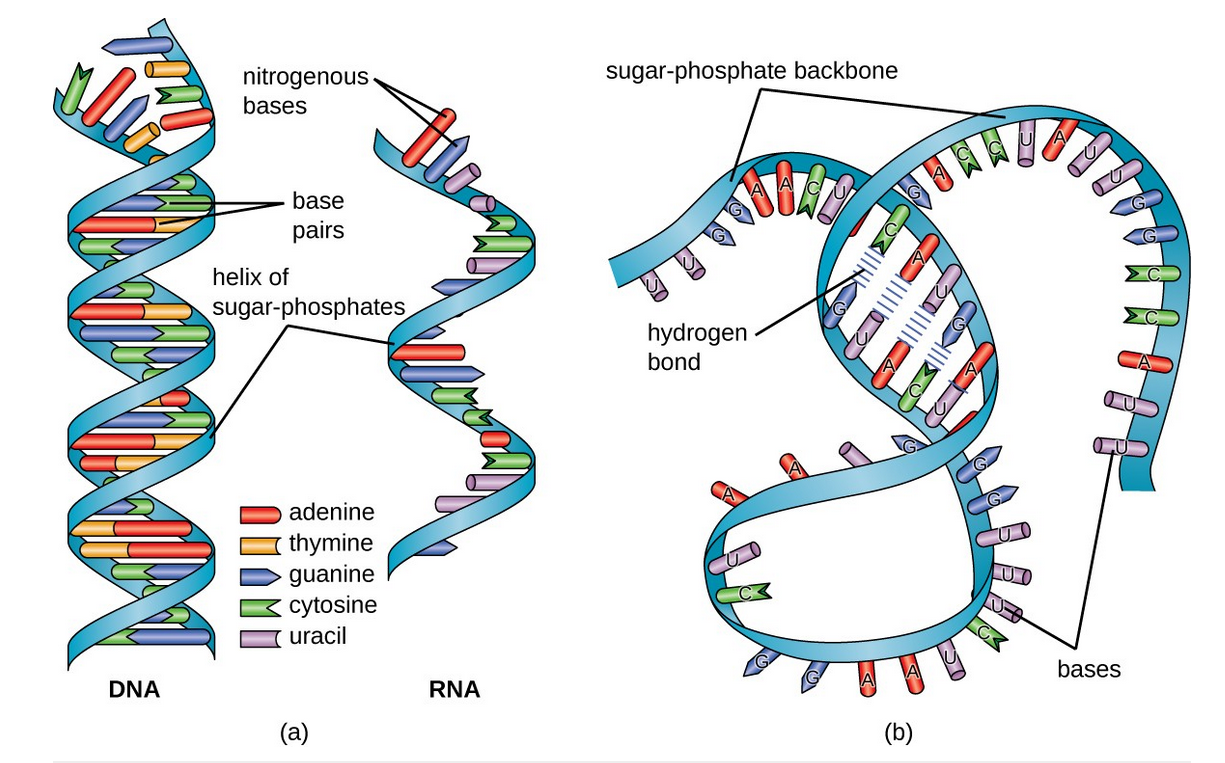
The aim of my study is to collect data on lncRNAs associated with cervical and lung cancer – mainly from the database lnc2cancer 3.0 and put these sequences through thorough analysis using bioinformatics tools. Putative G-quadruplex forming sequences can be predicted using web-based tools like QGRS Mapper and G4 hunter. Both these tools use algorithms to identify these G-rich sequences and give output regarding the G-tetrads found and their respective G-scores.

This scattered information – one from the lncRNA database and one from the bioinformatics tools can be collated into a very useful pre-experimental analysis which can be further used for in-vitro work if expertly curated. This is exactly what this study strives to do. A few select lncRNAs can be filtered out from this hybrid database created and taken to the next step of analysis. A deep dive into the expression pattern of these lncRNAs in-vitro can yield promising results in cancer research.

**REVIEW OF LITERATURE**

1. **RNA (Ribonucleic Acid):**

Nucleic acids are the building blocks of life and RNA is an essential part of cell processes. It was first put forward by Alexander Rich and Linus Pauling in 1953 using X-ray crystallography [5]. Ribonucleic acid is a complex, single stranded linear molecule which contains 4 nucleotides – Adenine (A), Cytosine (C), Uracil (U) and Guanine (G). Each of these bases consists a phosphate group, a ribose sugar and a nitrogenous base [6]. These nucleotides are bound to each other through phosphodiester bonds. RNA plays many essential roles in our cells like protein synthesis, switching genes on and off, transcription initiation, mRNA splicing, and translation [1].



**Figure 1 (a):** A comparison between the structure of DNA and RNA. **(b)** Structure of RNA which folds unto itself forming a three-dimensional structure.

*Source:* *Structure of a Ribonucleic Acid Marquisee, Susan H . Merrill, John Robert Penswick and Ada Zamir, American Association for the Advancement of Science.*

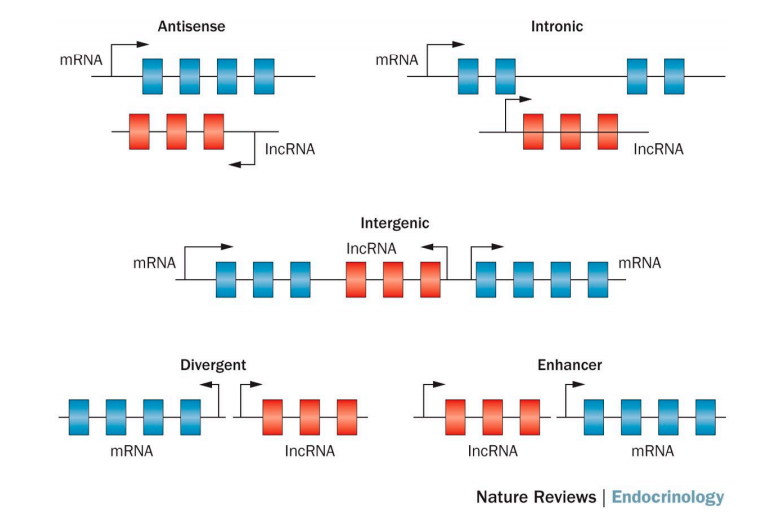
1. **Non-Coding RNAs:**

It is no secret that RNAs play a major role in cell processes and a ton of evidence has been collected in the past years which depicts that the non-coding part of the genome is of much more significance than previously understood. In June 2017, it was discovered that only 3.56% of the disease-associated SNPs (single nucleotide polymorphisms) were found in the protein-coding region and the rest 96.44% were found in the non-coding regions [7].

Non-coding RNA (ncRNA) is transcribed from the DNA molecule but not translated into a functional protein. In the past few years, there has been a substantial increase in the interest of these ncRNAs and more research is being conducted on them than ever before (Figure 3). There are basically two types of ncRNAs – small ncRNAs with less than 200 nucleotides and long ncRNAs with more than 200 nucleotides. Our focus lies on the latter.

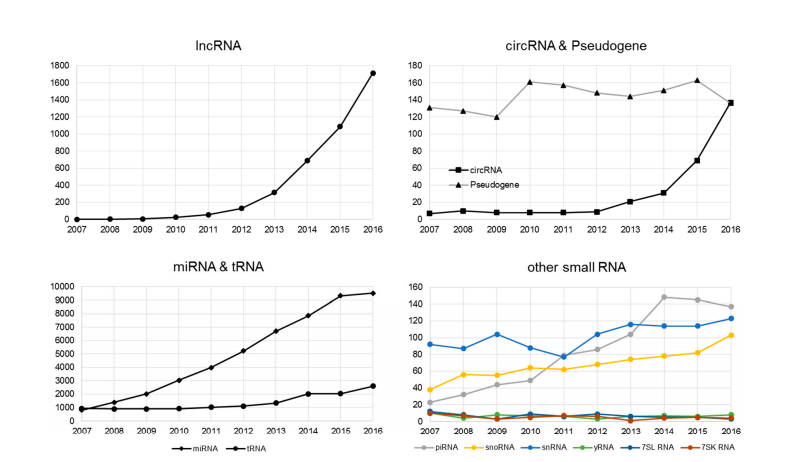
1. **Long Non-Coding RNAs:**

As mentioned above, lncRNAs are non-coding RNAs which are more than 200 nucleotides long. They were first discovered mainly through bioinformatics tools and analysis of transcriptome data. They are transcribed by RNA polymerase II and are segregated into different types – antisense, intronic, divergent, intergenic and enhancer according to their genome position (Figure 2). LncRNAs were always believed to be non-functional but recent studies have shown that they are significant to many processes like embryonic stem cell differentiation, high order chromosomal dynamics, telomere biology and subcellular structural organization. [8]



**Figure 2:** Biogenesis of long non-coding RNAs (lncRNAs)

*Source: Wang, J., Samuels, D. C., Zhao, S., Xiang, Y., Zhao, Y. Y., & Guo, Y. (****2017****). Current research on non-coding ribonucleic acid (RNA). Genes, 8(12)*



**Figure 3:** Increase in the research conducted on ncRNAs by Pubmed searches

*Source: Wang, J., Samuels, D. C., Zhao, S., Xiang, Y., Zhao, Y. Y., & Guo, Y. (****2017****). Current research on non-coding ribonucleic acid (RNA). Genes, 8(12)*

The major contribution of lncRNAs comes in the form of their role in cancer. In recent years, many lncRNAs are proved to have a role in certain cancers – be it due to upregulation or downregulation. For example, MALAT-1 was associated with poor prognosis in non-small cell lung cancer and oral squamous cell carcinoma. Similarly, lncRNA HOTAIR has been observed to promote metastasis in multiple cancers including but not limited to breast, colorectal, cervical and liver cancer [7]. LncRNAs can also be used as biomarkers for cancer – one example being the lncRNA ZEB1-AS1. It predicts unfavourable prognosis in gastric cancer and, lncRNA-ATB for hepatocellular carcinoma.

1. **LncRNAs and Cancer:**

LncRNAs are a diverse group of transcripts that regulate gene expression through a variety of mechanisms. As a result, they've been discovered to be highly expressed in tumours, and they've been connected to the transformation of healthy cells into tumour cells.

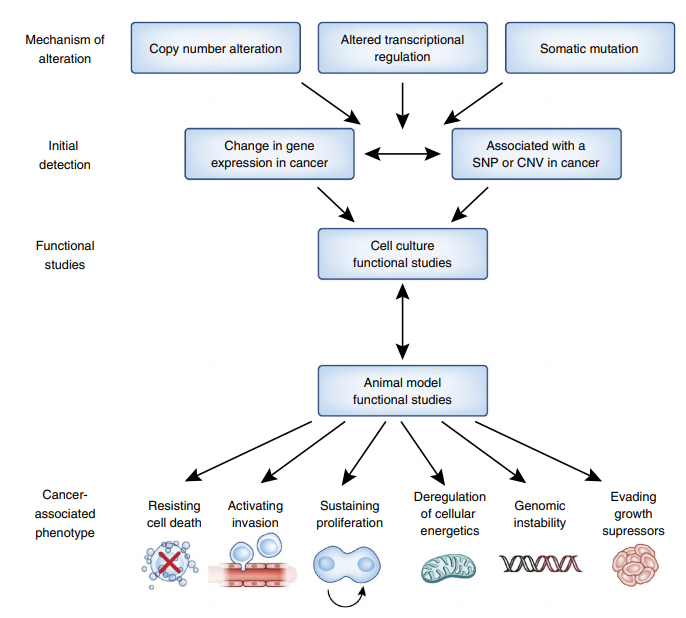
In terms of their role in cancer, lncRNAs have tissue-specific expression and are expressed in a controlled manner in correlation with distinct gene sets that influence cell cycle control, survival, immune response, or pluripotency, among several other functions, which decide the transformed phenotype of cancer cells.

|  |  |  |  |
| --- | --- | --- | --- |
| **lncRNA** | **Accession No** | **Cancer Phenotype** | **Cancer Association** |
| **GAS5** | Ensembl ID: ENSG00000234741 | Induces cell arrest and sensitizes cells to apoptosis. Alters cell metabolism | Downregulated in breast cancer |
| **LINC-PINT (MKLN1-AS1)** | GEO profile: FLJ43663 | Inhibits cell proliferation and promotes apoptosis | Inhibits cell proliferation and promotes apoptosis |
| **NBAT1 (CASC14)** | GEO profile: LOC729177 | Inhibits cell proliferation and invasion, and impairs differentiation of neuronal precursors | Inhibits cell proliferation and invasion, and impairs differentiation of neuronal precursors |
| **BCAR4** | Ensembl ID: ENSG00000262117 | Promotes proliferation and migration | Expression correlates with advanced breast cancer metastasis and anti-estrogen resistance |
| **PCAT1** | Ensembl ID: ENST00000561978.1 | Promotes cell proliferation | Upregulated in prostate cancer. Contains disease-associated SNPs |

**Table 1**: Examples of lncRNAs with tumour-suppressor or oncogenic functions

*Source: Huarte, M. (****2015****). The emerging role of lncRNAs in cancer. Nature Medicine, 21(11), 1253–1261.*

There a number of steps (Figure 4) which have to be considered while talking about lncRNAs in cancer. Firstly, there is identification of the lncRNAs in cancer. This is mainly done by deep-sequencing technologies and functional interrogation of lncRNAs. Then comes observing the alteration of lncRNA expression in cancer cells. Microarray technologies are usually used for this process using different types of cancer samples. Although, majority of the lncRNAs have been identified using high-thoroughput sequencing of RNA (RNA-seq). Furthermore, functional studies including in silico predictions have to be carried out for successful understanding. [7]

**Figure 4**: Identification and characterization of lncRNAs with roles in cancer.

*Source: Huarte, M. (****2015****). The emerging role of lncRNAs in cancer. Nature Medicine, 21(11), 1253–1261.*

**Cervical Cancer:**

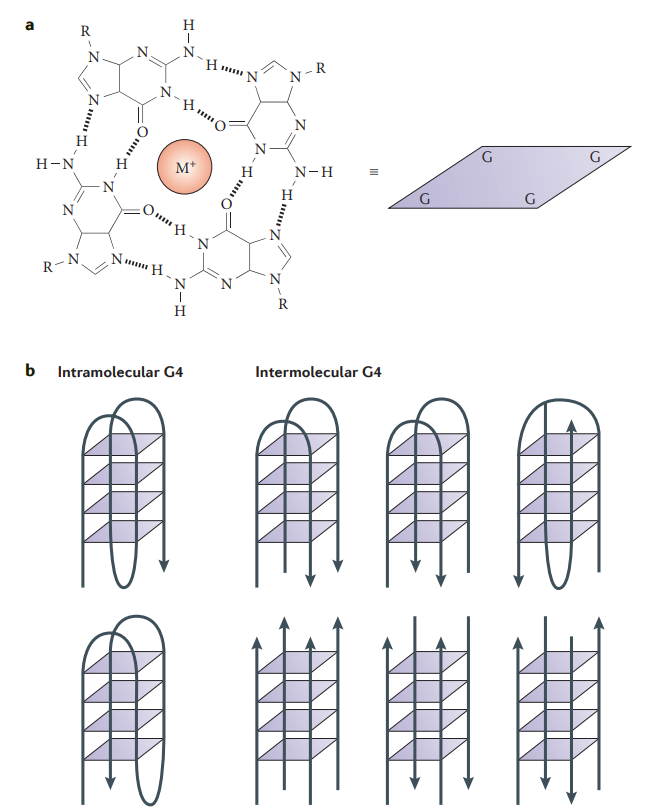
Cervical cancer is the world's second most common cause of cancer morbidity and mortality in women [9]. In their studies, Cao and colleagues have managed to prove that lncRNA GAS5 plays a significant role in tumorigenesis and cancer progression and the expression of GAS5 was significantly lower in cervical cancer cells in comparison to the surrounding cells. LncRNA ZFAS1 was shown to be up regulated in cervical cancer cells and silencing this lncRNA could possibly inhibit cell proliferation and cell migration [10].

**Lung Cancer:**

Lung cancer is one of the most common types of cancers in humans across the globe with possibly the highest mortality rate [11]. There is proven correlation between lncRNAs and select cancers but since this a relatively new field of research, not much data has been collected regarding specific cancers but there are a few notable lncRNAs. For example, lncRNA H19 is known to serve as a tumor suppressor gene in the pathogeneis of lung cancer. Other lncRNAs include HOTAIR, which might play a role in the early stages of lung cancer but further studies need to be conducted to give a sure function [11]. PVT1, another notable lncRNA, was noted to be up-regulated in lung cancer cells and higher expression correlated to lower survival rates in patients [12].

1. **G QUADRUPLEXES:**

We’re all aware about the conventional structure of the DNA double helix which was first discovered by James Watson and Francis Crick in 1951, but there are some interesting structures that are formed within this double helix one of which is of note to look at – the non-canonical structure of G-quadruplex. They are four-stranded secondary structures and are abundant in guanine-rich regions of the DNA. The structure is basically a stack of G-quartets – a cyclic planar arrangement of four Hoogsteen hydrogen bonded guanine residues. In 1910, it was first put forward through an experiment which showed high concentrations of guanylic acid form a gel, implying that G-rich sequences in DNA may form higher-order structures and then almost 50 years later it was properly identified through X-ray fiber diffraction studies [3]. Because of their high structural stability under physiological conditions and the broad distribution of sequences consistent with G4 formation, their study has gained a lot of momentum in recent years [1]. There is no one fixed topology associated with these simple but yet convoluted structures. There are multiple factors which affect these topologies like the glycosidic conformation (syn/anti), number of molecules of the nucleic acid and their formation (intramolecular, bimolecular or tetramolecular), the monovalent cation (they stabilize the structure in the order of K+>Na+>Li+), and the relative orientation of the strands [3].

****

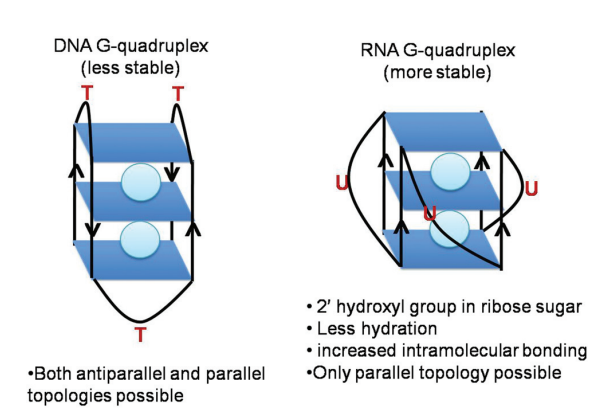
**Figure 5: (a)** Interactions in the G4 quartet. **(b)** Schematic of intermolecular and intramolecular structures of G4

*Source: Bochman, M. L., Paeschke, K., & Zakian, V. A. (2012). DNA secondary structures: Stability and function of G-quadruplex structures. Nature Reviews Genetics, 13(11), 770–780.*

1. **RNA G-QUADRUPLEX:**

While much research has been done on DNA G-quadruplexes, there has been a steady shift in recent years toward understanding G-quadruplexes at the RNA level. Various studies have come across the presence of RNA G4 in mRNA, long non-coding RNAs and telomeric ends [3]. Due to a multitude of factors, the presence of RNA G4 is much more likely than DNA. This is because DNA has its own pairing with the complementary strand, whereas RNA is single stranded and can form other secondary structures much more efficiently. Research done by Balasubramanian et al shows the presence of RNA G4 within the cytoplasm of human cells [13].

There are a few structural differences between the DNA and RNA G4. The first one comes with RNA having a ribose sugar and DNA having deoxyribose. The presence of 2’ hydroxyl groups in ribose sugar leads to a forceful conformation on the G4 which in turn results in added stability. This makes the RNA G-quadruplex’s topology independent of the environmental conditions [3].

****

**Figure 6**: Difference between DNA and RNA G-quadruplexes

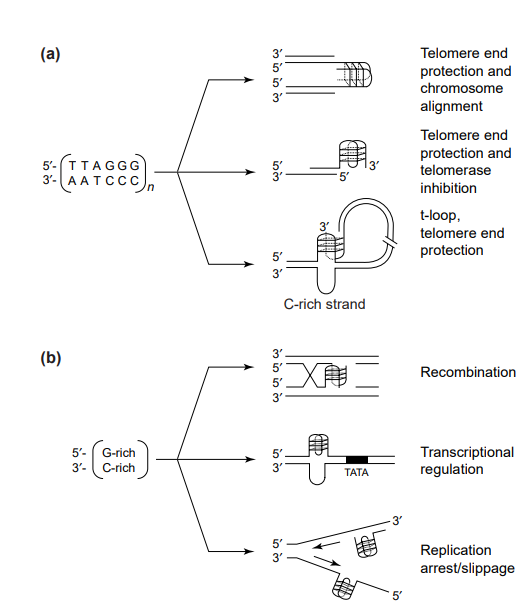
*Source: Agarwala, P., Pandey, S., & Maiti, S. (2015). The tale of RNA G-quadruplex. Organic and Biomolecular Chemistry, 13(20), 5570–5585.*

1. **G-QUADRUPLEX AS TARGETS FOR CANCER THERAPY:**

There are multiple ways in which G-quadruplexes can be used as targets for cancer therapy. This discussion begins with telomeres – a structure present at the ends of chromosomes. In cancer cells, telomerase (a protein complex which helps in the elongation of telomere sequence) activity is highly increased. In such cases, a telomere G4 can be used as a potential biomedical target which inhibits telomerase activity [2]

Recent studies have also shown that G-rich sequences which have the tendency to form G-quadruplex structures are present in promoter regions of a few important genes such as retinoblastoma susceptibility genes and the c-myc gene. This means that possibly G4 structures might play a role in the regulation of gene transcription [14]. This function of G4 can be explored and used as anti-cancer treatments further down the line.

One notable mention is the drug Quarfloxin. Quarfloxin, a ligand that interacts with G4, had completed Phase II trials as a candidate therapeutic agent against several tumors, including neuroendocrine tumors, carcinoid tumors, and lymphoma [15].



**Figure 7:** Schematic of the formation of G-quadruplex DNA during normal cellular events. (a) Guanine (G)-rich sequences of telomeres form different types of G-quadruplex structures at different stages of the cell cycle to protect the chromosome ends. In addition, G-quadruplex formation can align chromosomes and inhibit telomerase. (b) G-rich sequences in non-telomeric regions of the genome might transiently form G-quadruplex structures from duplex DNA to participate in cellular events such as recombination and transcription or to interfere with cellular events such as replication. Abbreviation: C, cytosine**.**

*Source: Han, H., & Hurley, L. H. (****2000****). G-quadruplex DNA: A potential target for anti-cancer drug design. Trends in Pharmacological Sciences, 21(4), 136–142.*

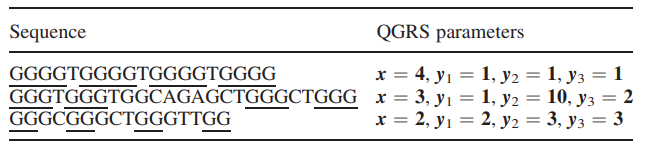
1. **QGRS Mapper:**

In-silico analysis of G-quadruplexes in the human genome has led to large-scale analysis of human genes. Since this a relatively new field of research, there are not many user-friendly and accessible tools which can be used to identify G-quadruplexes. This is where the web-based server QGRS Mapper steps in. This tool produces comprehensive information on the composition and distribution of putative quadruplex forming G-rich sequences (QGRS) through raw FASTA nucleotide sequences obtained from NCBI Nucleotide database. There are a number of options and filters that can be applied while searching for the G-quadruplexes – for example, minimum number of tetrads, maximum length of G4 motif and size as well as composition of the loops. G4 structures in oligonucleotides can also be predicted using this tool. This tool also maps Apart from giving out numerical data, QGRS mapper also provides a graphic representation of the data which can be useful for G4 visualization. [16]

The putative G4 are identified using the following motif:



Here, x is the number of guanine tetrads in the G4 and the length of the gap is denoted by y. There are a few restrictions in place like the sequence must contain at least two tetrads for it to be considered in the algorithm. Also, the user can choose a length between 30-45 bases. Additionally, users can also choose the loop or gap size between the G-groups.

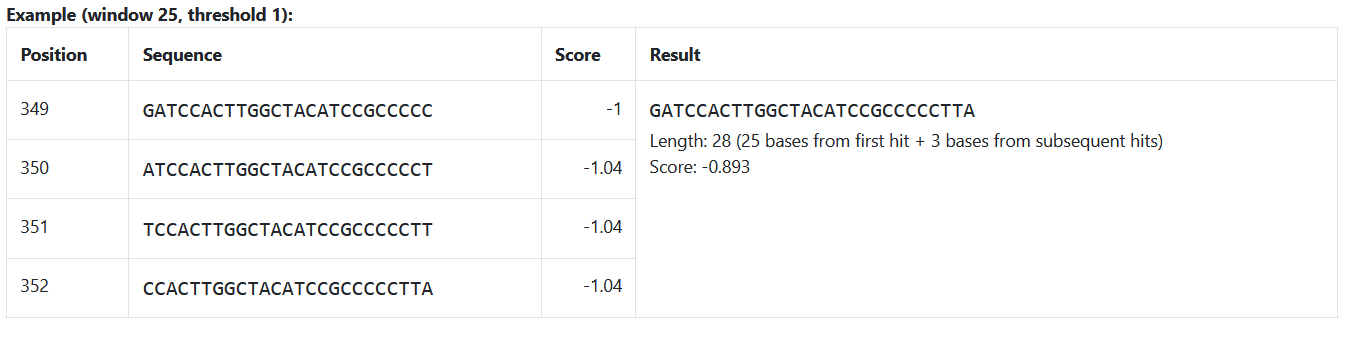
The table below depicts a few examples of a few QGRS which the algorithm recognizes:

*Source: Kikin, O., D’Antonio, L., & Bagga, P. S. (****2006****). QGRS Mapper: A web-based server for predicting G-quadruplexes in nucleotide sequences. Nucleic Acids Research, 34(WEB. SERV. ISS.), 676–682.*

The first sequence has four tetrads (4G) with equal gap length and the other two sequences have three tetrads or tracts (3G) and two tracts respectively (2G).

1. **G4 Hunter:**

There are a number of algorithms which recognize the G-quadruplex in a nucleotide sequence, but these algorithms mostly identify the local runs of G above a certain set threshold. Even with many filters available, there are a lot of deviations that arise with the other algorithms. Some sequences are false negatives which don’t show G4 formation in-silico but do form these quadruplexes in-vitro. Similarly, many sequences show false positives which means the opposite – they form G4 in-silico but not in-vitro. Additionally, a lot of these algorithms give us limited or more one-directional information on the G-quadruplexes formed – they basically don’t give a comprehensive quantitative analysis. In a bid to remedy this situation, the G4 hunter algorithm takes into account the G-richness (proportion of Gs in the sequence) and G-skewness (G/C asymmetry between the complementary strands) and scores accordingly. This algorithm was validated by using 209 sequences and observing G4 formation in-vitro. The calculation is done by assigning a score to each nucleotide – G bases have a positive score and C bases have a negative score. A and T have no scores associated with them [17]. Below is an example of the scores produced after a G4 hunter search result:



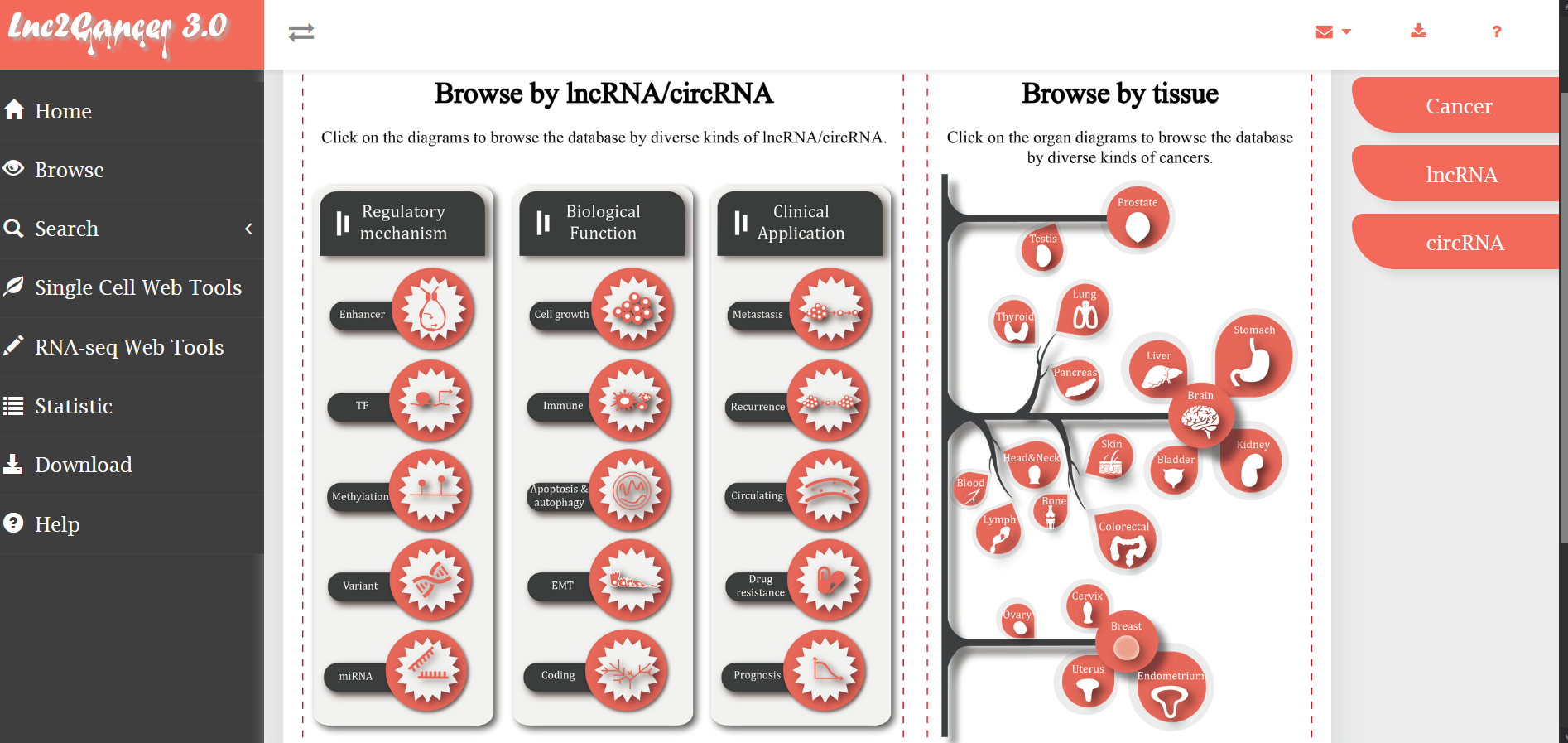
On the left most column is the position of the G-quadruplex in the nucleotide sequence and on the next columns are the nucleotide sequences, scores and the result respectively.

These scores from G4 hunter coupled with the data obtained from QGRS mapper ultimately provide us with substantial information on the lncRNAs for further experimental analysis.

**MATERIALS AND METHODS**

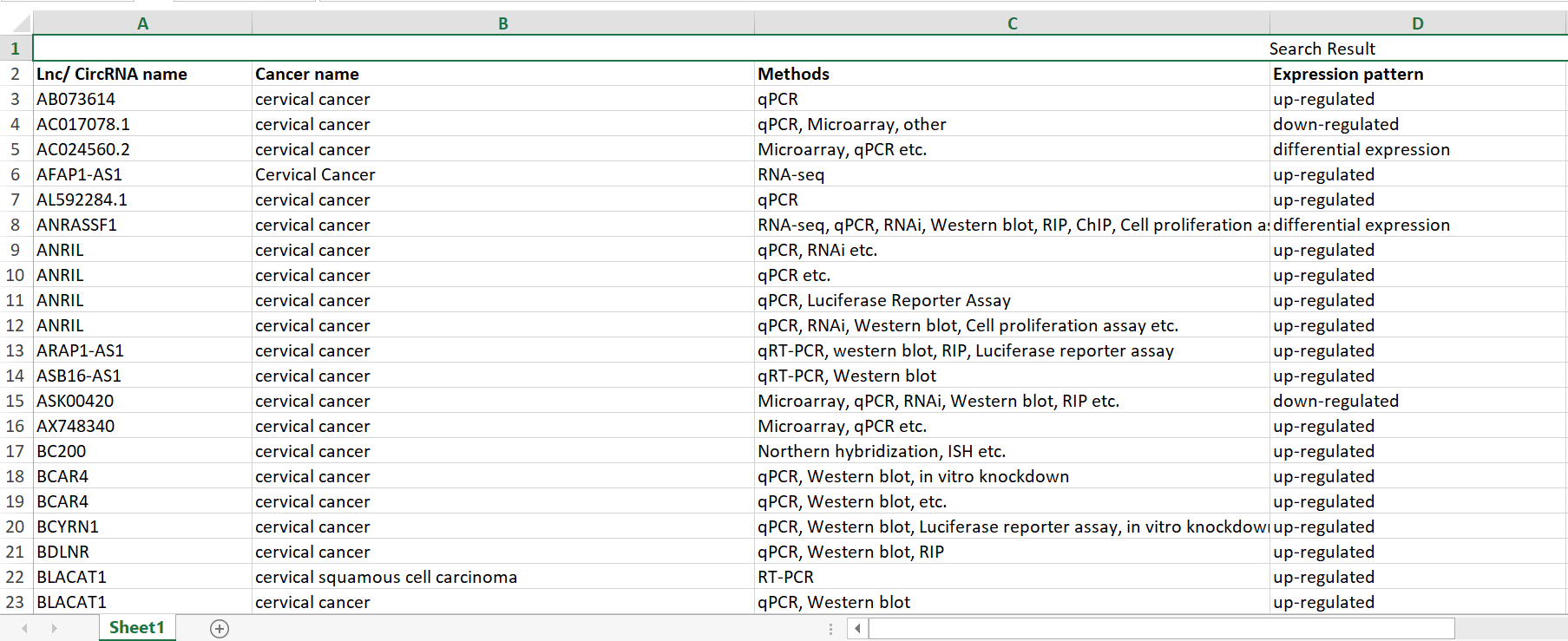
1. **Lnc2Cancer Database:**

The first step towards analysing lncRNAs was to find a reliable source of data. Lnc2Cancer 3.0 provided a comprehensive list of lncRNAs which are associated with the different human cancers [18]. The user simply has to select the type of cancer they wish and they are presented with the data to download. The Lnc2Cancer user interface is depicted below:



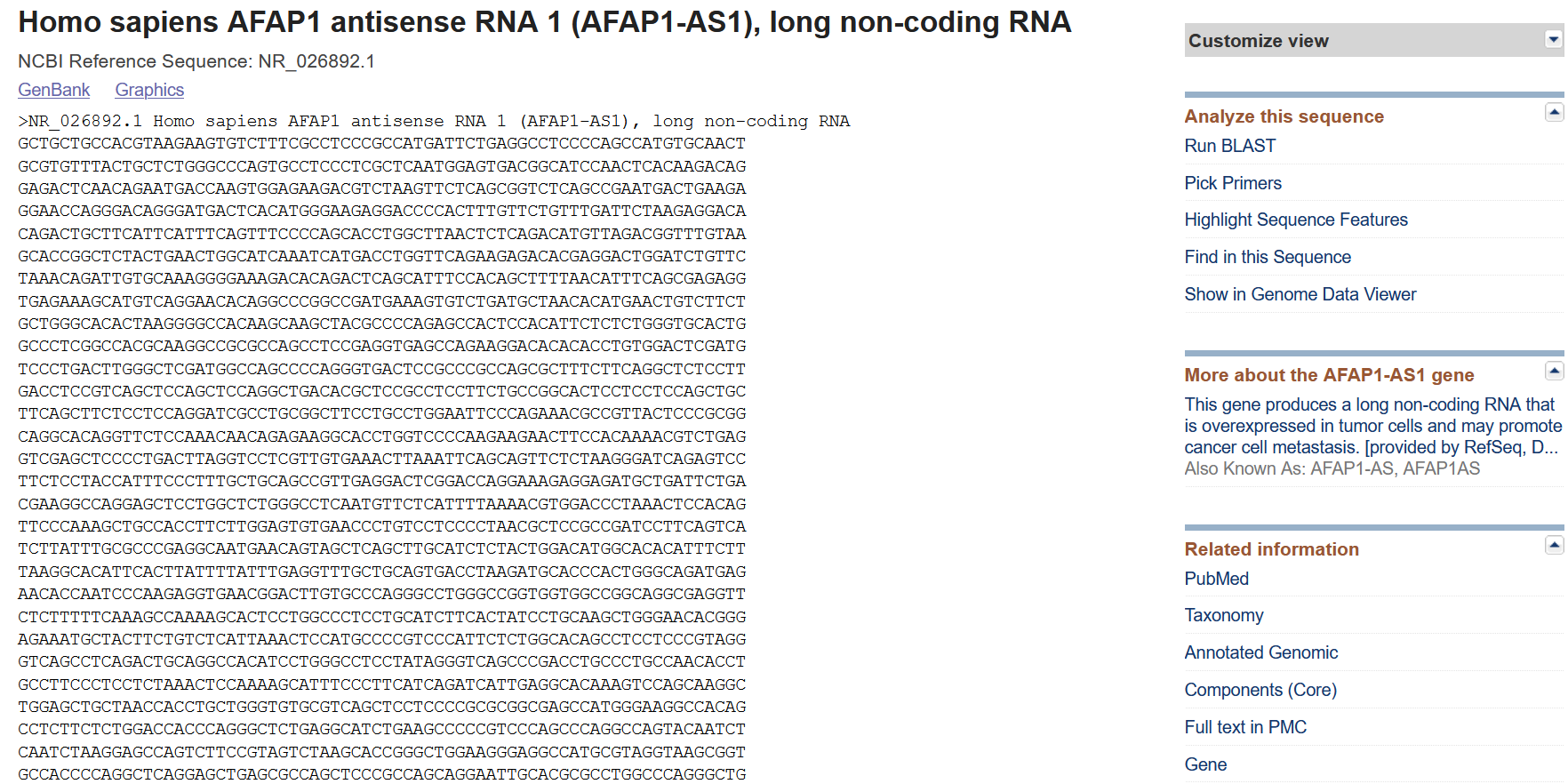
**Figure 8:** lnc2cancer 3.0 database user interface.

After selecting a particular cancer from the list above, we are provided with detailed information on all the lncRNAs associated with the cancer. The excel window below depicts the data retrieved from Lnc2cancer. There were 392 entries for cervical cancer. Information such as the methods of experimental data and the expression pattern of the lncRNA in the cancer. Further analysis of these sequences is done using QGRS Mapper.



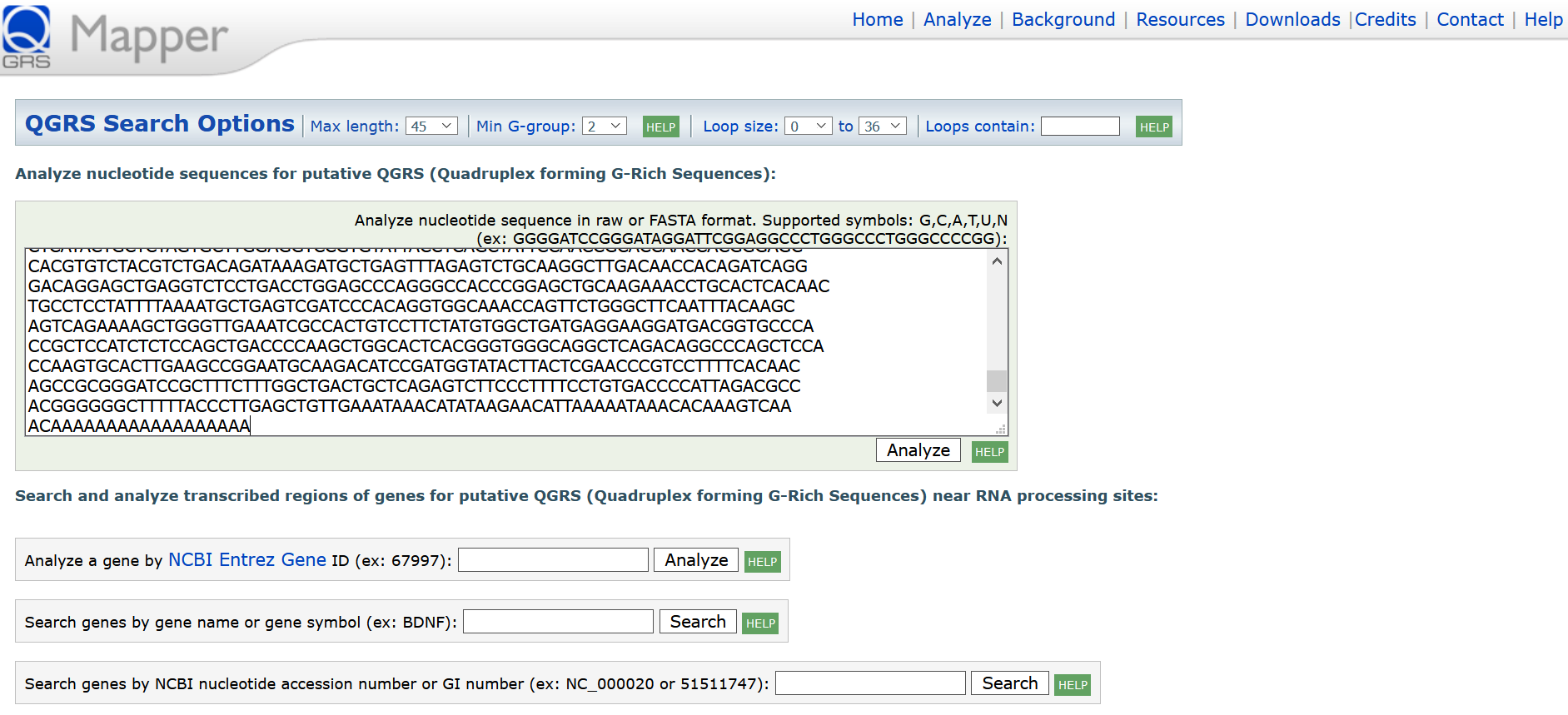
**Figure 9:** List of lncRNAs associated with cervical cancer obtained from lnc2cancer 3.0

1. **Using QGRS Mapper to analyse lncRNAs**:

* The analysis began by using this tool for predicting G-rich sequences which can form G-quadruplex. The first step was to retrieve the raw FASTA format from NCBI Nucleotide at <https://www.ncbi.nlm.nih.gov/nucleotide/>.

**Figure 10:** FASTA format of AFAP-AS1.

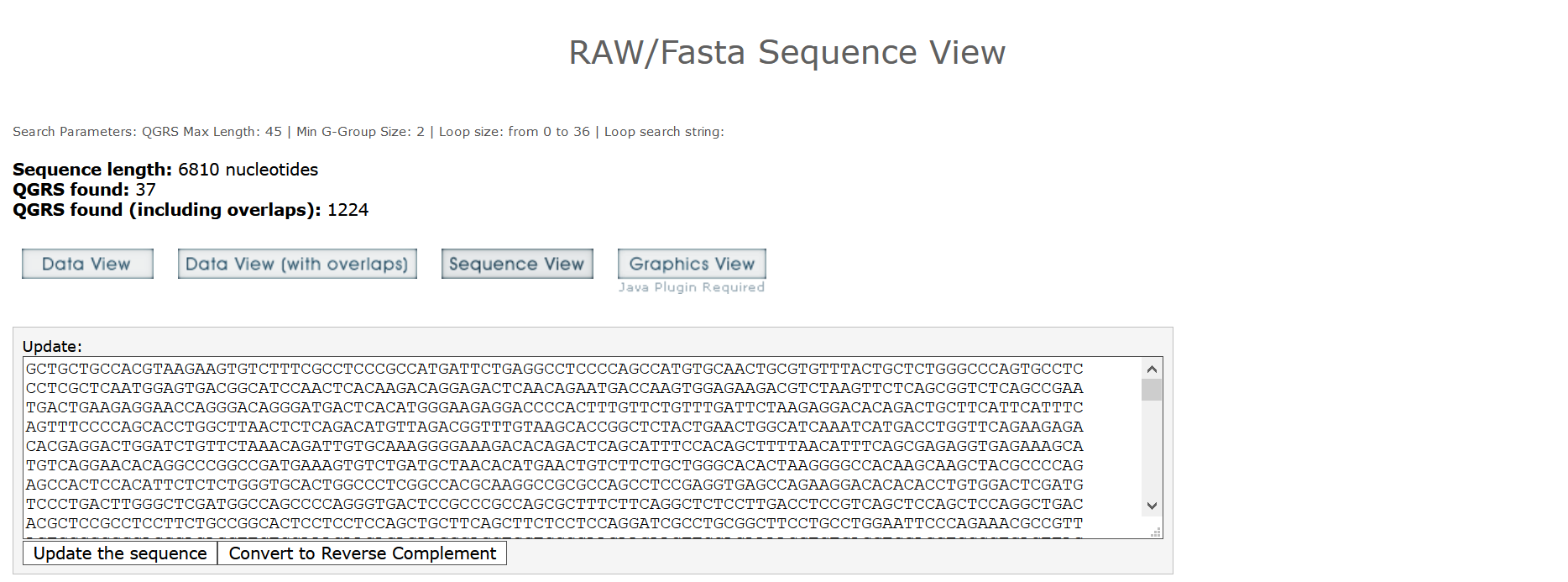
* Next, this raw FASTA format of the nucleotide sequence was input in the QGRS Mapper server (https://bioinformatics.ramapo.edu/QGRS/index.php):



**Figure 11:** QGRS Mapper User Interface

Search parameters; Max length of sequence: 45, Min G-group: 2, Loop size: 0 and Loop contain: blank.

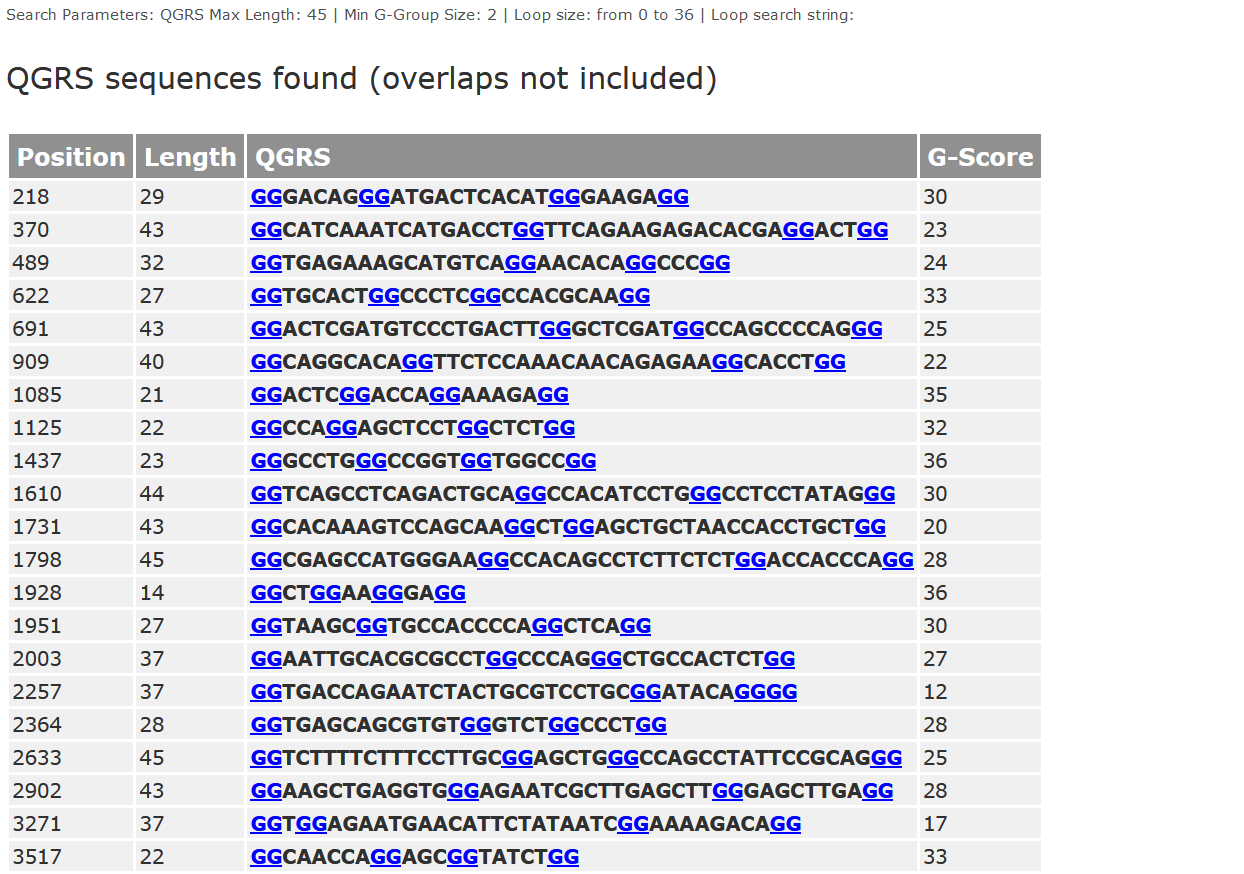
These search parameters were used for analysing all the sequences.

* The search yielded the number of quadruplex forming G-rich sequences. Since, the algorithm used the ‘sliding-window method’ for analysing the sequences, there are a few overlaps which are removed by the algorithm in the output by selecting the higher scoring G-quadruplexes [16].

**Figure 12:** Analysis view in QGRS Mapper.

The lncRNA AFAP1-AS1 showed 37 quadruplex forming sequences (without overlaps).

* More information about these sequences was found out by going to the ‘Data View’

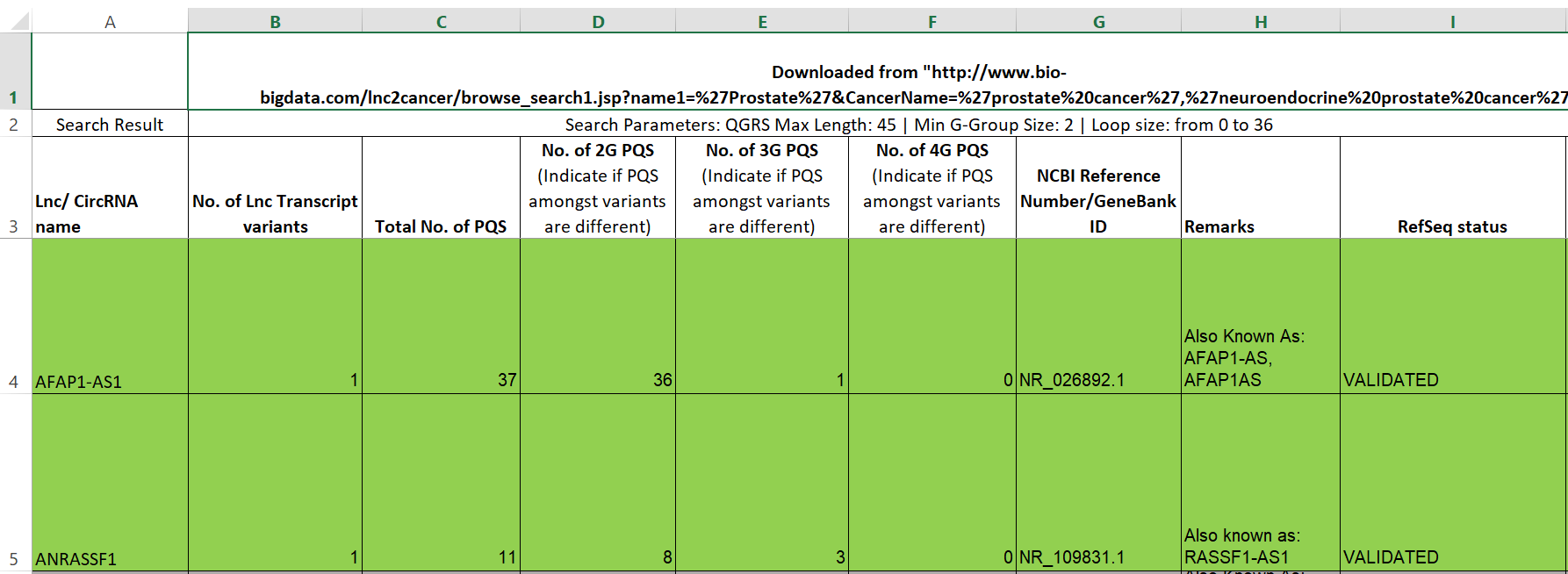


**Figure 13:** List of PQS shown in QGRS Mapper.

The left most column shows the starting position of the G4 forming sequence in the lncRNA, next the length of the sequence is depicted and the different QGRS with the G-score is also mentioned.

There are a few factors which determine the scores given to the G4-forming sequences:

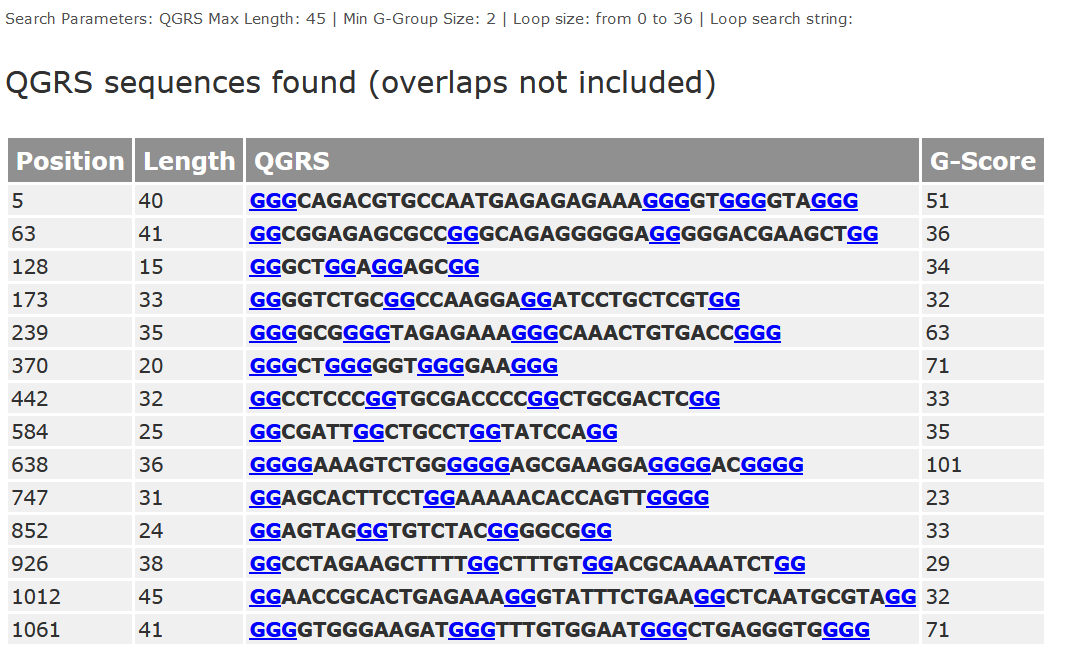
* The loop length, shorter loops confer more stability than longer loops.
* Most G4 have equally-sized loops.
* The number of guanine tetrads also determines the score. More the number of guanine tetrads, higher will be the G-score.
* All this information was collected and entered into the Excel database:



**Figure 14:** Excel datasheet created for information collected from QGRS Mapper.

The following columns were input in the excel sheet:

* LncRNA name: the name of the lncRNA obtained from the lnc2cancer 3.0 database.
* No. of lnc transcript variants: some lncRNAs has more than one transcript variants and each of the variants were individually assessed.
* No. of 2G PQS: total number of sequences containing 2 G-tracts were to be mentioned here (Figure)
* No. of 3G PQS: total number of sequences containing 3 G-tracts were to be mentioned here (Figure)
* No. of 4G PQS: total number of sequences containing 4 G-tracts were to be mentioned here (Figure)
* NCBI Reference number/GeneBank ID: this information was obtained from the NCBI Nucleotide website.
* Remarks: the alternate names for the lncRNA and any other observations were mentioned here.
* RefSeq Status: this information was also obtained from the NCBI Nucleotide website. The different input options were: Validated, Reviewed or Predicted.



**Figure 15:** List of PQS for lncRNA TOB1-AS1.

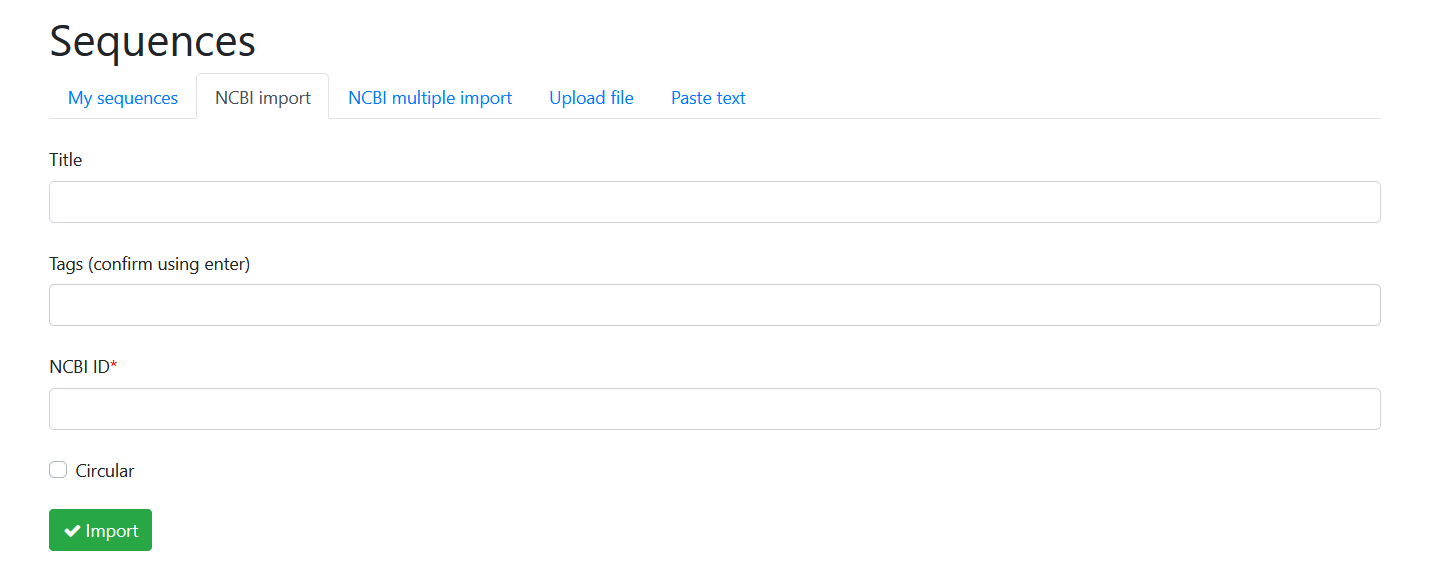
The above table is the analysis of lncRNA TOB1-AS1. The different 2G, 3G and 4G PQS are depicted here. The first quadruplex forming sequence begins at the 5th position in the nucleotide sequence and has 3 G-tetrads. Similarly, the second sequence forms 2 G-tetrads and the ninth sequence starts at the 638th position and forms 4 G-tetrads.

**A total of 701 sequences were analysed using QGRS Mapper – 192 for cervical cancer and 509 for lung cancer**.

1. **Analysing lncRNAs using G4Hunter:**

To narrow down our list of validated lncRNAs further, we used the web-based tool G4 hunter (<http://bioinformatics.ibp.cz/#/>).

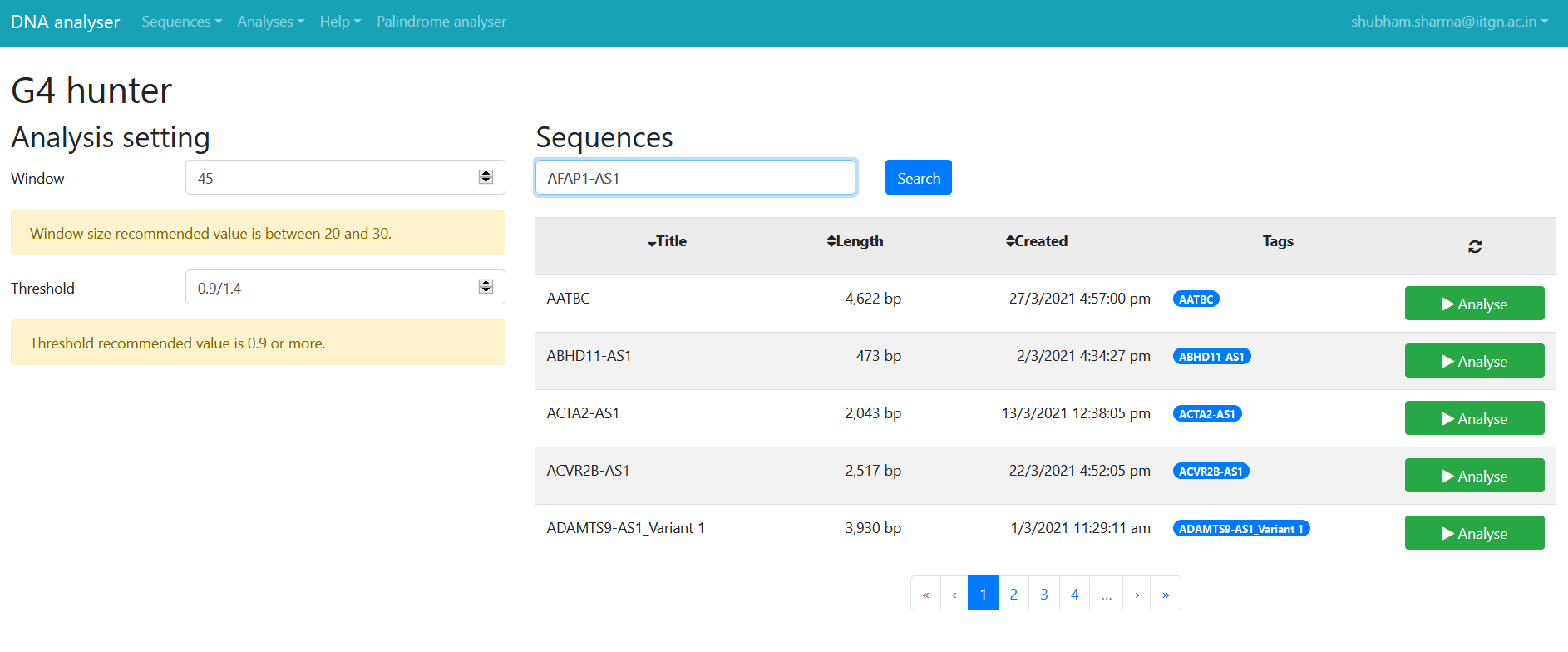
* First step was to upload the sequences to the G4 hunter server:



**Figure 16:** Importing sequences from NCBI Nucleotide to G4 Hunter.

Using the ‘NCBI Import’ option, the sequence details like title, tags and NCBI ID obtained from the NCBI Nucleotide page have to be uploaded. Once all the sequences were uploaded, we moved on to further analysis.

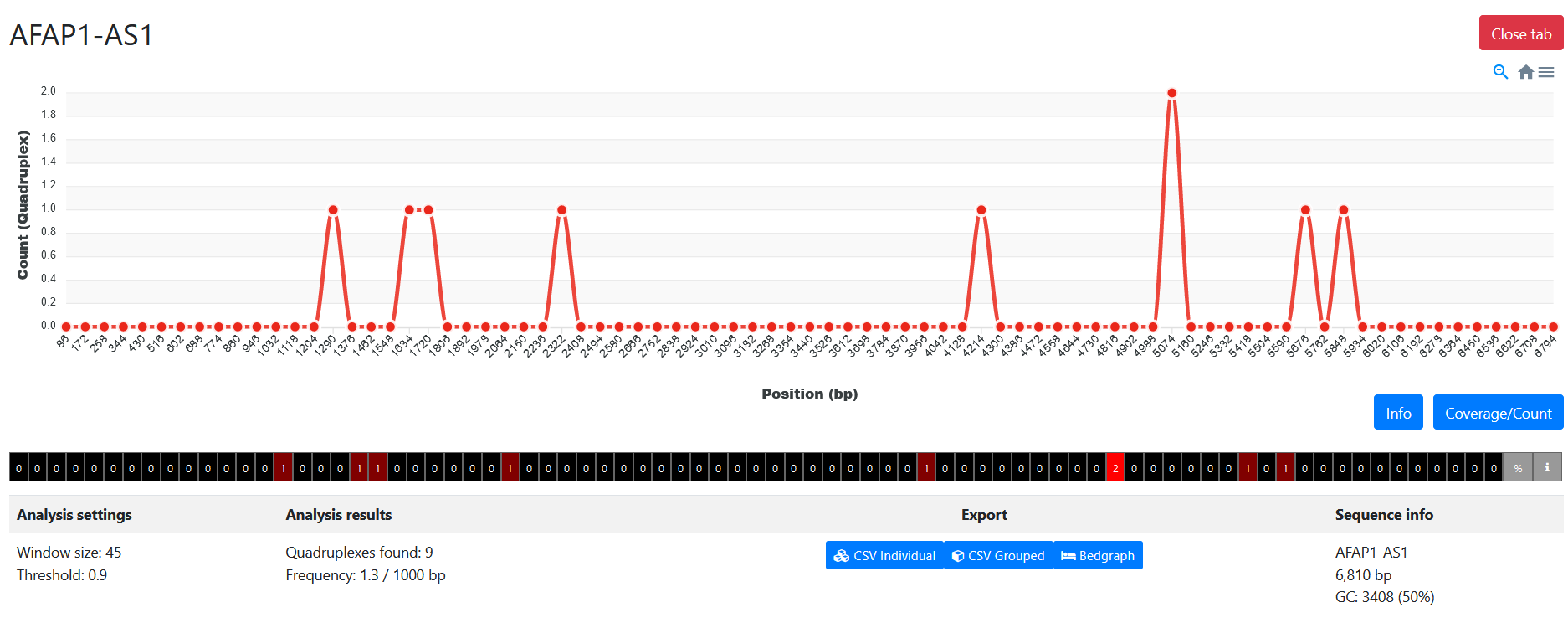
* Next, the parameters for the search and analysis have to be set, which can be carried out in this way:



**Figure 17:** Analysis interface of G4 Hunter.

The window size was selected as 45 (similar to QGRS Mapper) and the sequences were analysed in two thresholds –0.9 and 1.4. This was done to ensure that the PQS which score high (mostly 3G and 4G) are not lost at a lower threshold score – for example, a 4G PQS would score somewhere between 1.00-1.4, so to ensure that these 3G or 4G PQS are considered, each sequence was analysed in both threshold values.

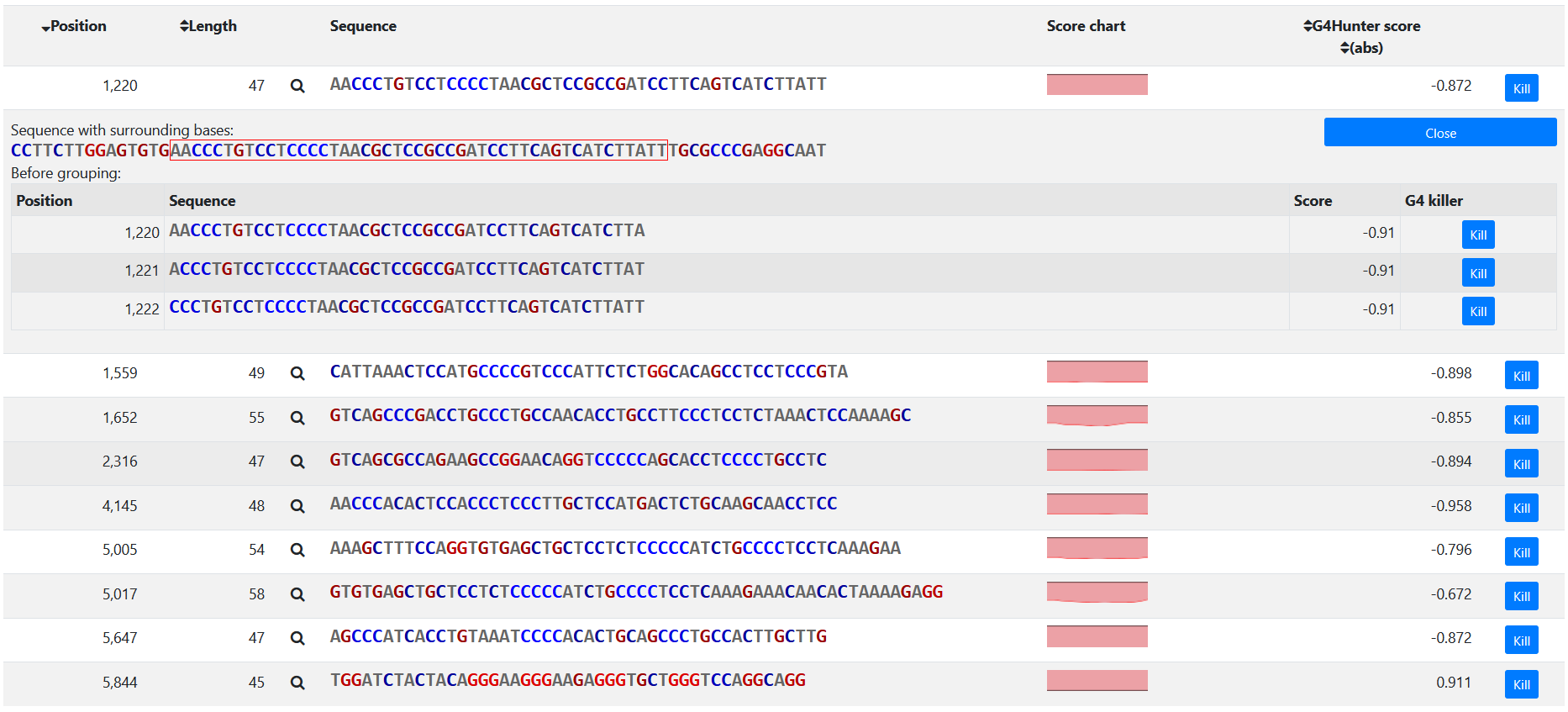
* The sequence is analysed and yields the following window:



**Figure 18:** G4 Hunter Analysis of lncRNA AFAP1-AS1.

Using the example AFAP1-AS1 (same as QGRS Mapper), G4 hunter predicts that there are 9 PQS in this sequence as opposed to the prediction of QGRS Mapper which was 37. The following information is portrayed in the analysis:

* Window size: the length of the sequence which is analysed
* Threshold: the baseline score, if the sequences have a score less than this then they won’t be displayed.
* Quadruplexes found: the number of putative quadruplex forming sequences found
* Frequency: The number of G4 formed per 1000 bp
* Sequence Information: the title, number of base pairs and GC content (%) is displayed
* Furthermore, the information on the sequences found with their details and overlaps is also provided:



**Figure 19:** List of G-quadruplexes predicted in G4 Hunter.

Here, the blue sequences are C-rich scored negatively according to the scoring algorithm and can form G-quadruplexes in their complementary strands. Red colour and a positive score is depicted for the G-rich sequences which can potentially form G-quadruplex. Other information such as starting positions of the sequence and the length of the sequence is also provided.

**A total of 448 sequences were analysed in G4 hunter: 141 for cervical cancer and 307 for lung cancer.**

**RESULTS AND DISCUSSION**

1. **Quantification of data obtained from QGRS Mapper:**

* **Cervical Cancer:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Type of PQS in 141 LncRNAs** | | | |
|  | **Total Validated Sequences** | **Sequences  having 2G PQS** | **Sequences  having 3G PQS** | **Sequences  having 4G PQS** |
| **Total no. of LncRNAs having PQS** | 141 | 141 | 58 | 6 |
| **Percentage** |  | 100.00% | 41.13% | 4.26% |

Out of a total of 189 entries, 141 lncRNAs had a ‘Validated’ or ‘Reviewed’ RefSeq status and were further quantified:

**Table 2:** Information on the types of PQS in cervical cancer lncRNAs and their percentage occurrence**.**

All of the validated sequences had 2G PQS since it is pretty common to find two G-tracts as compared to 3G and 4G. The least found PQS were of 4G nature.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **G-Score Range in 6 4G LncRNAs** | | | | | |
|  | **Total** | **86-90** | **91-95** | **96-100** | **101-105** | **106-110** |
| **Total No. of 4G LncRNAs having G-Scores** | 6 | 1 | 0 | 0 | 2 | 3 |

**Table 3:** Information on the G-scores of lncRNAs having 4G PQS

The G-scores of lncRNAs having 4G PQS was observed to be higher than the rest. The six lncRNAs mentioned are: DDN-AS1, DLG1-AS1, H19, STXBP-5, TOB1-AS1 and WTI-AS.

* **Lung Cancer:**

Out of a total of 509 entries, 307 lncRNAs had a ‘Validated’ or ‘Reviewed’ RefSeq status and were further quantified:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Type of PQS in 307 LncRNAs** | | | |
|  | **Total Validated Sequences** | **Sequences  having 2G PQS** | **Sequences  having 3G PQS** | **Sequences  having 4G PQS** |
| **Total no. of LncRNAs having PQS** | 307 | 307 | 121 | 8 |
| **Percentage** |  | 100.00% | 39.41% | 2.61% |

**Table 4:** Information on the types of PQS in lung cancer lncRNAs and their percentage occurrence**.**

Similar to cervical cancer, all of the validated sequences had 2G PQS and the number of lncRNAs having 3G and 4G were much less comparatively.

Separate data was collected on the 4G PQS:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **G-Score Range in 8 4G LncRNAs** | | | | | |
|  | **Total** | **86-90** | **91-95** | **96-100** | **101-105** | **106-110** |
| **Total No. of 4G LncRNAs having G-Scores** | 8 | 2 | 0 | 1 | 2 | 3 |

**Table 5:** Information on the G-scores of lncRNAs having 4G PQS

As observed before, the G-scores of lncRNAs having 4G PQS was much higher than 2G or 3G PQS. The 8 lncRNAs having 4G PQS are: CEBPA-AS1, GASL1, H19, IRAIN, KCNQ1OT1, LINC00336, LINC01124 and SLC9A3-AS1.

1. **Quantification of data obtained from G4 hunter:**

**Figure 20:** Bar graph depicting the PQS distribution for cervical cancer lncRNAs

Different patterns are seen for both 0.9 and 1.4 G-score thresholds. LncRNAs with 3G and 4G PQS show up in 1.4 threshold due to their high G-scores. A few lncRNAs having 4G PQS also show up in the 0.9 threshold due to the G4 hunter algorithm. The algorithm using ‘sliding window’ action to analyse the sequences and the putative G-quadruplex forming sequences are considered with their overlaps and an average score is taken out from all the overlaps, hence this discrepancy is observed.

The lncRNAs having 4G PQS in G4 hunter analysis for cervical cancer are as follows (without overlaps): DDN-AS1, DLG1-AS1, STXBP5-AS1, TOB1-AS1 and WTI-AS.

**Figure 21:** Bar graph depicting the PQS distribution for lung cancer lncRNAs

A similar pattern for lung cancer is also observed. A few lncRNAs with 2G PQS also show up in 1.4 threshold owing to the process of the G4 hunter scoring algorithm mentioned in the above cervical cancer analysis. LncRNAs with 4G PQS are as follows (without overlaps): CEBPA-AS1, EPB41L4A-AS2, FBXL19-AS1, GASL1, GATA2-AS1, IRAIN, LINC00336, LINC00982, LINC01124 and SLC9A3-AS1.

1. **Discussion:**

In depth literature review has shown that sequences which form G-quadruplexes with longer G-tracts are more stable. Hence, the lncRNAs which have predicted 4G PQS in both QGRS Mapper and G4 Hunter results can be picked out. These lncRNAs produced high G-scores in both the web-based tools and can be further targeted for in-vitro analysis. A few examples of such lncRNAs are:

|  |  |  |
| --- | --- | --- |
| lncRNA name | G score in QGRS Mapper (out of a possible ~158) | G score in G4 Hunter |
| DDN-AS1 | 106 | 0.873 |
| GASL1 | 106 | 1.125 |
| STXBP5-AS1 | 106 | 0.733 |
| SLC9A3-AS1 | 89 | 1.155 |
| IRAIN | 104 | 1.424 |

**Table 6:** G-scores for five lncRNAs in QGRS Mapper and G4 Hunter

A few studies conducted on these lncRNAs show that these RNA sequences do have a major role to play in cervical and lung cancer and can be potential drug targets for treatment.

* DDN-ASI:

This lncRNA was found to be overexpressed in cervical cancer. TCF3, a transcription factor for DDN-AS1 was also discovered which also regulates its expression [19]. This oncogenic function of TCF3 was found to be regulated by miRNAs and a few candidate miRNAs were selected using various bioinformatics tools [20]. Among these, miR-15a and miR-16 were demonstrated to have an anti-cancer effect on certain cancers. DDN-AS1 competitively binds to these miRNAs by acting as ceRNAs and hence upregulating these transcription factor TCF3 [21]. In this case, G-quadruplexes formed in lncRNA DDN-AS1 can be targeted to inhibit binding of the lncRNA to any miRNA.

* GASL1:

Growth-arrest associated lncRNA or GASL1 is found to be downregulated in non-small cell lung cancer and inhibits cell proliferation by possibly down-regulating TGF-beta1 [22]. The serum GASL1 can prove to be a sensitive biomarker for diagnosis and prognosis of lung cancer but an efficient use for it has not been developed due to ineffective measures. It is possible to target the G-quadruplexes formed in GASL1 for successful experimentation.

* STXBP5-AS1:

This lncRNA was found to have a reduced expression in cervical cancer cells. The overexpression of STXBP5-AS1 resulted in decreased cell proliferation by suppressing miR-96-5p expression. Similar to DDN-AS1, this lncRNA might have a possible role to play by acting as ceRNA and regulating the miRNA expression [23].

* SLC9A3-AS1:

The expression of this lncRNA is significantly increased in lung cancer cells and can be used as a potential biomarker for early diagnosis of lung cancer. Studies have been done where lncRNAs are extracted from extracellular vesicles (EV) present in peripheral blood. The methods used to identify these lncRNAs need to be extremely sensitive since there is not much quantity of lncRNAs found in EVs. In this regard, the G-quadruplexes can play a role in developing a better diagnostic measure.

* IRAIN:

This lncRNA was found to promote tumorigenesis in non-small cell lung cancer. In vitro studies proved that the knockdown on this lncRNA led to significantly suppressed cell proliferation [24]. The mechanism of action of this lncRNA remains unclear, but developing a drug which targets the G-quadruplexes through small molecule ligands can prove to be an effective way to decrease tumour proliferation in lung cancer patients.

This data coupled with the studies done on the G-quadruplexes can be applied to experimentation and can be further taken to in-vitro analysis.

**CONCLUSION**

The studies conducted on G-quadruplexes and lncRNAs are pretty recent when we pit them against tried and tested methods for cancer therapy, and the research done by correlating both these fields is even less prevalent. Multiple examples of candidate lncRNAs which can be used for cancer treatment were listed, but this is just speculation based on literature review. The creation of this database using multiple sources and trying to collate that information is just a small first step in a long thought-out journey. Further examination of these lncRNAs needs to be done in the form of experimental analysis, to at the most proving their prospective significance in other biological processes, apart from cancer treatments. The studies done with lncRNA cancer database, QGRS Mapper, G4 hunter and lncATLAS can be further explored through coupling them with various other bioinformatics tools, and then possibly this has the potential to become the one of the most prominent measures to treat various human ailments.

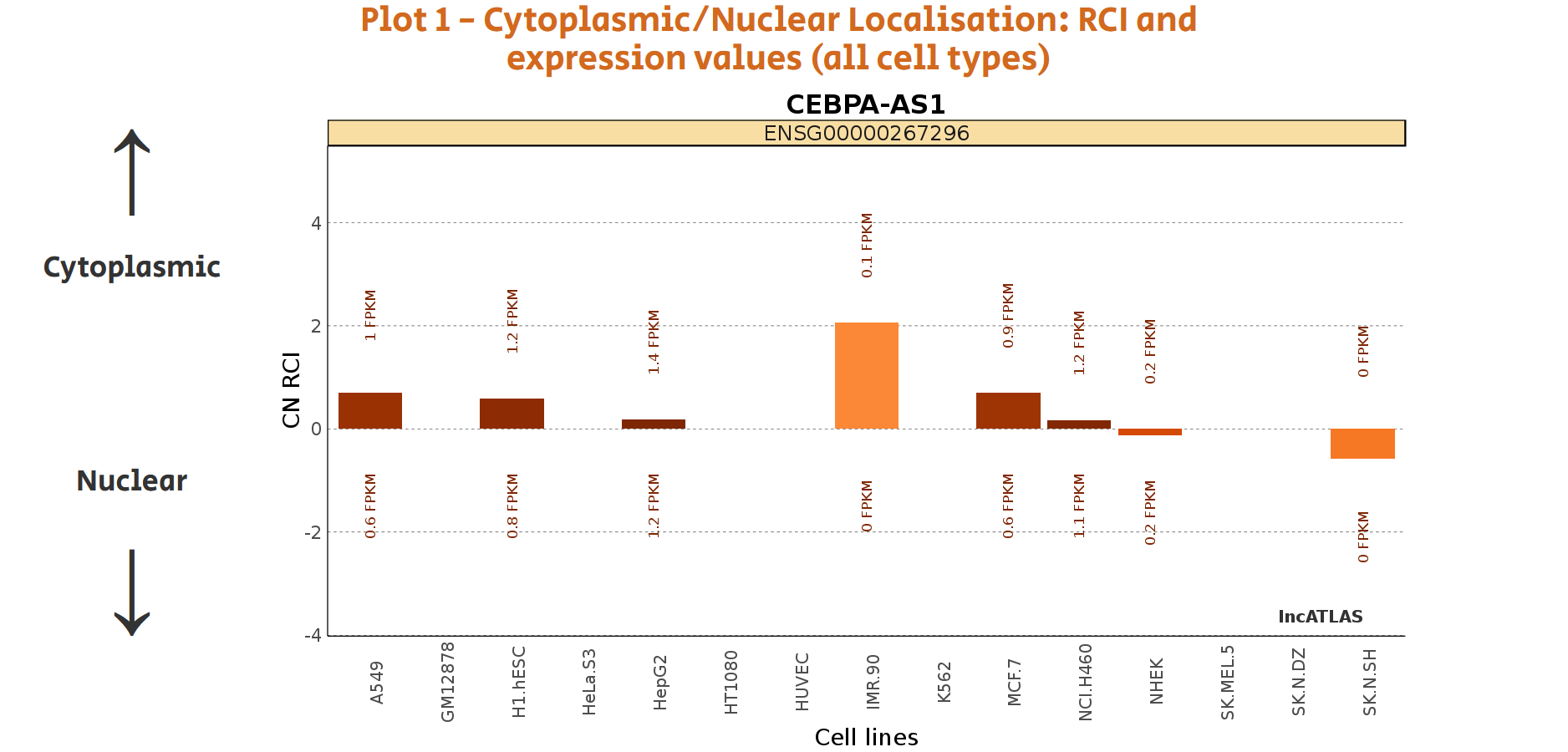
**FUTURE PROSPECTS**

Long non-coding RNAs are believed to perform a range of diverse functions in the cell, but they are poorly conserved at the sequence level which might make functional annotation tough to gauge [25]. The functionality of lncRNAs depends on appropriate subcellular localization. The lncRNA transcripts are observed to be localized in many different places within the cell like chromatin, nucleus, cytoplasm and exosomes [26].

Information on subcellular localization can possibly paint a clearer picture about the exact biological function of the lncRNA [27].

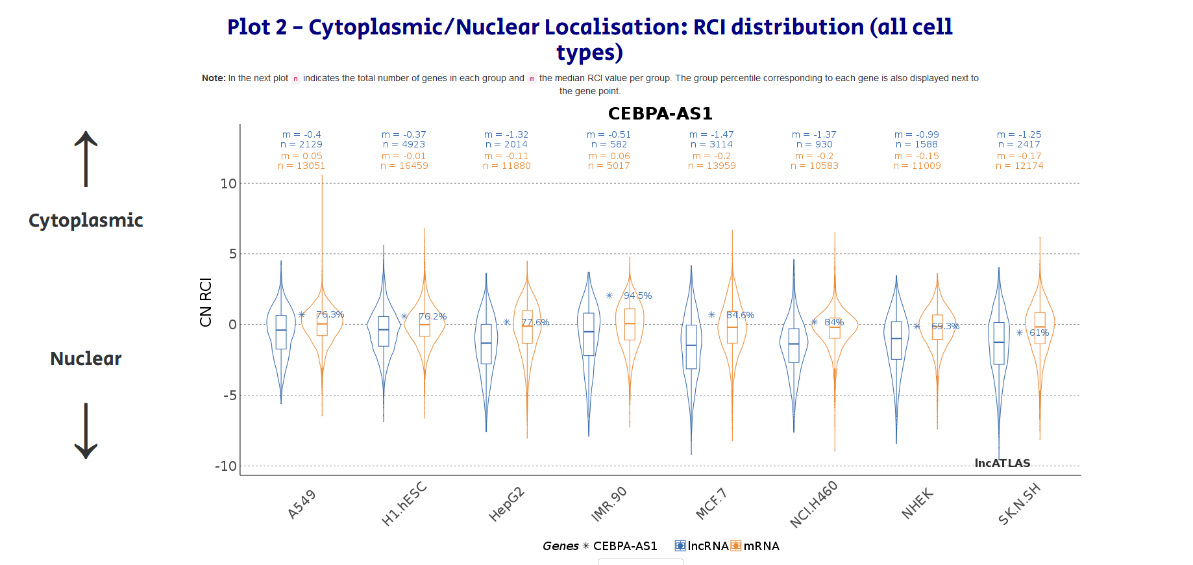
It might be experimentally challenging and expensive to figure out localization in newly found lncRNAs, and this is where the database lncATLAS comes into play. It gives information on lncRNA localization in human cells based on RNA-sequencing datasets [28]. The data for creating this bioinformatics tool was collected from the largest dataset of subcRNAseq, produced by the ENCODE consortium and raw RNAseq data from a panel of human cell lines were used to quantify the reference GENCODE gene annotation [29]. An example of lncRNA CEBPA-AS1 subcellular localization from lncATLAS is mentioned below:

* CEBPA-AS1:



**Figure 22:** Plot 1 for subcellular localization in CEBPA-AS1

To make things more practical, a relative scheme to portray subcellular localization was adopted, with RCI being the ‘Relative Concentration Index’. RCI is the ratio of a transcript’s concentration, per unit mass of sampled RNA, between 2 two compartments (nucleus and cytoplasm) [28]. The x-axis shows the different cell types and the bars are coloured respective to the expression levels of the lncRNA in the two compartments. FPKM denotes fragments per kilo base per million mapped and CN-RCI denotes cytoplasmic/nuclear RCI.



**Figure 23:** Plot 2 for subcellular localization in CEBPA-AS1

To put RCI values in context, their percentile rank within the distribution of all lncRNAs is indicated (ranks 12 relative to lowest value). The plot shows CN-RCI values distribution of all lncRNAs (blue) and mRNAs (orange) for each cell line (“n” indicates total number of genes, “m” median of CN-RCI values for lncRNAs and mRNAs separately). Some reference plots are also depicted using lncRNAs like DANCR (cytoplasm localized) and MALAT1 (nuclear localized).

LncRNA subcellular localization using effective bioinformatics tools is a sure fire way to ensure optimal experimentation of lncRNAs in-vitro. This is just one of many ways to take this study of lncRNAs and G-quadruplexes forward.

**REFERENCES**

1. Bochman, M. L., Paeschke, K., & Zakian, V. A. (**2012**). DNA secondary structures: Stability and function of G-quadruplex structures. *Nature Reviews Genetics*, *13*(11), 770–780.

2. Asamitsu, S., Obata, S., Yu, Z., Bando, T., & Sugiyama, H. (**2019**). Recent progress of targeted G-quadruplex-preferred ligands toward cancer therapy. *Molecules*, *24*(3).

3. Agarwala, P., Pandey, S., & Maiti, S. (**2015**). The tale of RNA G-quadruplex. *Organic and Biomolecular Chemistry*, *13*(20), 5570–5585.

4. Zhang, R., Xia, L. Q., Lu, W. W., Zhang, J., & Zhu, J. S. (**2016**). LncRNAs and cancer. *Oncology Letters*, *12*(2), 1233–1239.

5. Rao, S., & Arora, K. (**2020**). Recent trends in molecular techniques for food pathogen detection. In *Chemical Analysis of Food* (pp. 177–285). Elsevier.

6. Holley, R. W., Apgar, J., Everett, G. A., & Madison, J. T. (**2017**). Structure of a Ribonucleic Acid Marquisee , Susan H . Merrill , John Robert Penswick and Ada Zamir Published by : American Association for the Advancement of Science Stable URL : http://www.jstor.org/stable/1715055 REFERENCES Linked references are availab, *147*(3664), 1462–1465.

7. Huarte, M. (**2015**). The emerging role of lncRNAs in cancer. *Nature Medicine*, *21*(11), 1253–1261.

8. Wang, J., Samuels, D. C., Zhao, S., Xiang, Y., Zhao, Y. Y., & Guo, Y. (**2017**). Current research on non-coding ribonucleic acid (RNA). *Genes*, *8*(12).

9. Cao, S., Liu, W., Li, F., Zhao, W., & Qin, C. (**2014**). Decreased expression of lncRNA GAS5 predicts a poor prognosis in cervical cancer. *International Journal of Clinical and Experimental Pathology*, *7*(10), 6776–6783. Retrieved from www.ijcep.com/

10. Feng, L. L., Shen, F. R., Zhou, J. H., & Chen, Y. G. (**2019**). Expression of the lncRNA ZFAS1 in cervical cancer and its correlation with prognosis and chemosensitivity. *Gene*, *696*(December 2018), 105–112.

11. Yang, J., Lin, J., Liu, T., Chen, T., Pan, S., Huang, W., & Li, S. (**2014**). Analysis of lncRNA expression profiles in non-small cell lung cancers (NSCLC) and their clinical subtypes. *Lung Cancer*, *85*(2), 110–115.

12. Yang, Y. R., Zang, S. Z., Zhong, C. L., Li, Y. X., Zhao, S. S., & Feng, X. J. (**2014**). Increased expression of the lncRNA PVT1 promotes tumorigenesis in non-small cell lung cancer. *International Journal of Clinical and Experimental Pathology*, *7*(10), 6929–6935.

13. Biffi, G., Tannahill, D., McCafferty, J., & Balasubramanian, S. (**2013**). Quantitative visualization of DNA G-quadruplex structures in human cells. *Nature Chemistry*, *5*(3), 182–186.

14. Han, H., & Hurley, L. H. (**2000**). G-quadruplex DNA: A potential target for anti-cancer drug design. *Trends in Pharmacological Sciences*, *21*(4), 136–142.

15. Drygin, D., Siddiqui-Jain, A., O’Brien, S., Schwaebe, M., Lin, A., Bliesath, J., … Rice, W. G. (**2009**). Anticancer activity of CX-3543: A direct inhibitor of rRNA biogenesis. *Cancer Research*, *69*(19), 7653–7661.

16. Kikin, O., D’Antonio, L., & Bagga, P. S. (**2006**). QGRS Mapper: A web-based server for predicting G-quadruplexes in nucleotide sequences. *Nucleic Acids Research*, *34*(WEB. SERV. ISS.), 676–682.

17. Bedrat, A., Lacroix, L., & Mergny, J. L. (**2016**). Re-evaluation of G-quadruplex propensity with G4Hunter. *Nucleic Acids Research*, *44*(4), 1746–1759.

18. Gao, Y., Shang, S., Guo, S., Li, X., Zhou, H., Liu, H., … Zhang, Y. (**2021**). Lnc2Cancer 3.0: An updated resource for experimentally supported lncRNA/circRNA cancer associations and web tools based on RNA-seq and scRNA-seq data. *Nucleic Acids Research*, *49*(D1), D1251–D1258.

19. Liu, Z., Wu, M., Shi, H., Huang, C., Luo, S., & Song, X. (**2019**). DDN-AS1-miR-15a/16-TCF3 feedback loop regulates tumor progression in cervical cancer. *Journal of Cellular Biochemistry*, *120*(6), 10228–10238.

20. Li, H., Sheng, Y., Zhang, Y., Gao, N., Deng, X., & Sheng, X. (**2017**). MicroRNA-138 is a potential biomarker and tumor suppressor in human cervical carcinoma by reversely correlated with TCF3 gene. *Gynecologic Oncology*, *145*(3), 569–576.

21. Li, Z., Jiang, P., Li, J., Peng, M., Zhao, X., Zhang, X., … Li, X. (**2018**). Tumor-derived exosomal lnc-Sox2ot promotes EMT and stemness by acting as a ceRNA in pancreatic ductal adenocarcinoma. *Oncogene*, *37*(28), 3822–3838.

22. Su, W. Z., & Yuan, X. (**2018**). LncRNA GASL1 inhibits tumor growth of non-small cell lung cancer by inactivating TGF-γ pathway. *European Review for Medical and Pharmacological Sciences*, *22*(21), 7282–7288.

23. Shao, S., Wang, C., Wang, S., Zhang, H., & Zhang, Y. (**2019**). LncRNA STXBP5-AS1 suppressed cervical cancer progression via targeting miR-96-5p/PTEN axis. *Biomedicine and Pharmacotherapy*, *117*(April), 109082.

24. Feng, J., Sun, Y., Zhang, E. B., Lu, X. Y., Jin, S. D., & Guo, R. H. (**2015**). A novel long noncoding RNA IRAIN regulates cell proliferation in non small cell lung cancer. *International Journal of Clinical and Experimental Pathology*, *8*(10), 12268–12275.

25. Geisler, S., & Coller, J. (**2013**, October 9). RNA in unexpected places: Long non-coding RNA functions in diverse cellular contexts. *Nature Reviews Molecular Cell Biology*. Nature Publishing Group.

26. Hewson, C., & Morris, K. V. (**2016**). Form and Function of Exosome-Associated Long Non-coding RNAs in Cancer. *Current topics in microbiology and immunology*.

27. Gudenas, B. L., & Wang, L. (**2018**). Prediction of LncRNA Subcellular Localization with Deep Learning from Sequence Features. *Scientific Reports*, *8*(1), 1–10.

28. Mas-Ponte, D., Carlevaro-Fita, J., Palumbo, E., Hermoso, T., Guigo, R., & Johnson, R. (**2017**). LncATLAS database for subcellular localisation of long noncoding RNAs. *bioRxiv*.

29. Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., … Snyder, M. (**2012**, September 1). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Research*. Cold Spring Harbor Laboratory Press.